UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS

FACULTAD DE CIENCIAS MATEMATICAS

E.A.P. DE ESTADISTICA

Comparación mediante simulación de los métodos em e imputación múltiple para datos faltantes

TESIS

Para optar el Título Profesional de Licenciada en Estadística

AUTOR

Lourdes Angélica Galarza Guerrero

Lima – Perú

2013

COMPARACIÓN MEDIANTE SIMULACIÓN DE LOS MÉTODOS EM E IMPUTACIÓN MÚLTIPLE PARA DATOS FALTANTES

Lourdes Angelica Galarza Guerrero

Tesis presentada a consideración del Cuerpo Docente de la Facultad de Ciencias Matemáticas, de la Universidad Nacional Mayor de San Marcos, como parte de los requisitos para obtener el Título Profesional de Licenciada en Estadística.

Aprobada por:

.....
Lic. Grabiela Montes Quintana
(Presidenta)

.....
Lic. Rosa Fatima Medina Merino
(Miembro)

.....
Dr. Erwin Kraenau Espinal
(Miembro - Asesor)

Lima – Perú
Diciembre – 2013

FICHA CATALOGRÁFICA

GALARZA GUERRERO, LOURDES ANGELICA

Comparación mediante simulación de los métodos EM e

Imputación Múltiple para datos faltantes, (Lima) 2013.

vii, 83 p., 29.7 cm, (UNMSM, Licenciada, Estadística, 2013).

Tesis, Universidad Nacional Mayor de San Marcos,

Facultad de Ciencias Matemáticas 1. Estadística I.

UNMSM / F. de C.M. II. Título (Serie).

DEDICATORIA

A mis padres Felix y Luz,

que son mis pies sobre la tierra.

A mi hermana Rocio,

que son los brazos sobre los cuales

me apoyo y me impulso con más fuerza.

A mi mamá Elsa,

que se llevó mi corazón y

esta promesa hoy cumplida.

AGRADECIMIENTOS

La elaboración de este trabajo de investigación demandó la cooperación y consejería de algunas personas con la que tuve la especial suerte de contar y por las que expreso mi total agradecimiento:

A mis padres, por todo su apoyo incondicional en mi formación profesional.

A la Mg. Ana María Cárdenas Rojas, por su constante motivación y orientación en la elaboración de los pilares de mi trabajo de investigación.

A mi asesor Dr. Erwin Kraenau Espinal, por brindarme su confianza, sus consejos y en especial sus conocimientos que fueron de gran aporte para que el presente trabajo de investigación se volviera realidad.

A las autoridades y docentes de la Escuela Académico Profesional de Estadística – UNMSM, por brindarme la formación en esta maravillosa carrera profesional.

RESUMEN

COMPARACIÓN MEDIANTE SIMULACIÓN DE LOS MÉTODOS EM E IMPUTACIÓN MÚLTIPLE PARA DATOS FALTANTES

LOURDES ANGELICA GALARZA GUERRERO

DICIEMBRE - 2013

Orientador **: Dr. Erwin Kraenau Espinal**

Título Obtenido **: Licenciada en Estadística**

En el siguiente trabajo se presentan dos tratamientos a los problemas suscitados en el análisis de datos con presencia de datos perdidos: El Algoritmo EM basado en la Estimación por Máxima Verosimilitud y la Imputación Múltiple para datos faltantes, ambos métodos presentan ciertas ventajas frente a los métodos de imputación simple que ocasionan la obtención de estimadores distorsionados y sesgados. El algoritmo EM y la Imputación Múltiple se aplican a un conjunto de datos obtenido por simulación, causándole la pérdida de algunos valores con el objetivo de realizar posteriores comparaciones de las estimaciones obtenidas en casos con el conjunto de datos con y sin información faltante.

PALABRAS CLAVES: ALGORITMO EM

IMPUTACIÓN MÚLTIPLE

DATOS FALTANTES

MECANISMO DE DATOS FALTANTES

ABSTRACT

COMPARISON BY SIMULATION OF EM AND MULTIPLE IMPUTATION METHODS FOR DATA MISSING

LOURDES ANGELICA GALARZA GUERRERO

DECEMBER - 2013

Tutor **: Dr. Erwin Kraenau Espinal**

Academic Degree **: Licenciada en Estadística**

Two treatments to the issues raised in the analysis of data in the presence of missing data are presented in the following work: The EM algorithm based on Maximum Likelihood Estimation and Multiple Imputation for missing data, both methods have certain advantages over simple imputation that cause obtaining distorted and biased estimators. The EM algorithm and multiple imputation applied to a data set obtained by simulation, causing the loss of some values in order to make further comparisons of the estimates obtained in cases with dataset without missing information.

KEYWORDS: EM ALGORITHM

MULTIPLE IMPUTATION

MISSING DATA

MISSING DATA MECHANISM

ÍNDICE

CAPÍTULO III: ESTIMACIÓN POR MÁXIMA VEROSIMILITUD

CAPITULO IV: IMPUTACIÓN MÚLTIPLE DE DATOS FALTANTES

CAPÍTULO V: APLICACIONES

CAPÍTULO I MARCO REFERENCIAL

1.1 INTRODUCCIÓN

En estadística tratamos con una gran cantidad de datos, los cuales se consideran parte importante de todo trabajo de investigación, una de las situaciones recurrentes en cuanto al manejo de un conjunto de datos es la presencia de información faltante, es que el disponer de una base de datos completa es lo ideal para todo usuario o investigador sin embargo, la realidad difiere mucho con este ideal.

Los datos faltantes se consideran como aquella información que por algún motivo no pudo ser obtenida en el trabajo de campo, situación que puede perturbar el análisis estadístico de los datos debido a la disminución del tamaño muestral; afectando directamente la representatividad y la potencia de las pruebas estadísticas que se apliquen. A pesar que los usuarios tienen conocimiento de la presencia de datos faltantes, la mayoría trata el conjunto de datos como si éste estuviera completo, a veces haciendo uso de paquetes estadísticos que asumen de manera predeterminada un conjunto de datos completo, en estos casos los usuarios ignoran las consecuencias estadísticas que contraerían el hacer pasar desapercibido este problema.

Es así que, muy a pesar que los investigadores enfoquen sus esfuerzos en evitar la presencia de datos faltantes en su conjunto de datos, ésta es una situación con la que siempre tendrán que lidiar, especialmente cuando se trate de información sensible por parte del entrevistado o

unidad informante, allí surge la atención por parte de la comunidad investigadora en obtener métodos para el tratamiento de los datos faltantes.

A partir de la segunda mitad de la década de los ochenta del siglo XX los esfuerzos se centraron en entender las causas de este fenómeno e introducir una metodología para tratar los datos faltantes. De esta manera surgieron los métodos de imputación, técnica encargada de asignar valores reemplazantes a los valores faltantes en un conjunto de datos tomando como referencia información completa de la misma variable o de otras variables. Es así que la imputación es inicialmente presentada como una alternativa para manejar la omisión de los casos con información faltante.

Dentro del mismo contexto empezaron a surgir diversas metodologías que hacen uso de la imputación de datos de acuerdo al tipo y al comportamiento de las variables presentes en su análisis, las primeras metodologías utilizaban imputaciones consideradas simples debido a la asignación de un único valor por cada dato faltante y que por su fácil aplicación rápidamente formó parte de muchas bibliografías y trabajos de investigación muy a pesar que presentaba desventajas como la distorsión de los parámetros estimados.

La innovación y avances tecnológicos computacionales de la época ayudaron en el surgimiento de nuevas metodologías buscadas por los investigadores con el fin de mejorar los métodos de imputación simple, así es que se plantean dos nuevos e importantes métodos: el método de Máxima Verosimilitud (1970) y la imputación múltiple de datos (Rubin 1987),

métodos que consisten en buscar o asignar valores cercanos a la realidad de cada valor faltante y de esta manera obtener las estimaciones.

En el siguiente trabajo de tesis, el objetivo principal de investigación es la aplicación del algoritmo EM y la imputación múltiple ejecutándolas en el programa MATLAB y utilizando la simulación para la creación de un conjunto de datos con la distribución de probabilidad solicitada de acuerdo al método. Una vez realizada las imputaciones y obtenidas las estimaciones de los parámetros con ambas metodologías, se procede a realizar comparaciones de estos resultados y entre las estimaciones obtenidas de la muestra original de datos simulado.

Por tanto, el siguiente trabajo se ha dividido en cinco capítulos; en el primer capítulo se presenta el marco referencial con el que ha sido desarrollado la tesis; el segundo capítulo hace referencia a una revisión bibliográfica sobre los métodos de imputación simple; el tercer y cuarto capítulo se enfoca en el desarrollo del método de estimación por máxima verosimilitud con el algoritmo EM y la imputación múltiple de datos faltantes presentándose el diseño metodológico y el método de aplicación en ambos casos; finalmente en el quinto capítulo se presentan aplicaciones de ambas metodologías.

1.2 **IDENTIFICACIÓN DEL PROBLEMA**

La presencia de datos faltantes en un conjunto de datos ha sido y es considerado hasta el día de hoy como un contratiempo en el manejo de información, presentándose en la mayoría de encuestas como por ejemplo en las de hogares aplicados en distintos ámbitos

socioeconómicos y países. Las tasas de no respuesta parcial reflejan que muy a pesar que se intensifiquen medidas para evitar la ausencia de datos, hay casos en las que debido a la delicadeza de la información que se pide al entrevistado, éste muestra su negativa a responder y si lo hace sus respuestas no son acorde con su realidad.

Según la división de Estadística y Proyecciones Económicas CEPAL entre los años de 1990 – 1997 en los países de Latinoamérica se presenciaron casos de tasa de no respuesta en especial con variables relacionadas al ingreso en sus encuestas de hogares. Para el país argentino la tasa de no respuesta fue de 9,61 por ciento en 1997, para 1996 en Chile un 7,14 por ciento de asalariados no respondieron preguntas de ingreso, Colombia en el año 1997 tuvo como tasa de no respuesta un 8,61 por ciento mientras que en el mismo año en Costa Rica y Venezuela un 11,03 y 6,92 por ciento no declaraban ingresos en sus encuestas.

En el Perú cifras similares fueron obtenidas en el año 2008, la ENAHO registró información faltante que se reflejó en las tasas de no respuesta parcial con un 3,1 por ciento, para el año 2005 la tasa se incrementó a un 12,3 por ciento, mientras que con respecto a la información solicitada de los ingresos; un 25,9 por ciento de los entrevistados del quinto quintal (categoría con el ingreso más alto) se mostraron reacios a brindar dicha información.

Ante todo esto y como solución a la incidente tasa de no respuesta parcial la imputación se presentó como una metodología ventajosa frente a los inconvenientes presentados en los casos en que se omitía los faltantes. La Imputación de Medias, Imputación Cold Deck, Hot Deck y sus derivados, la imputación por Regresión y la Imputación haciendo uso de la

Estimación por Máxima Verosimilitud, son los métodos que hasta en la actualidad están siendo utilizados por investigadores internacionales y en el aspecto nacional en el Perú el Instituto Nacional de Estadística e Informática INEI ha utilizado estas herramientas en la Encuesta Nacional de Hogares ENAHO, haciendo que los sesgos producto de las omisiones de información sean tratados con procedimientos de imputación como el método Hot Deck para el caso de variables cualitativas y el método de Matrices Promedios para el caso de las variables cuantitativas (el ingreso).

Hasta la actualidad dichas metodologías son utilizadas para lidiar con este problema, a pesar de haberse hecho conocido que en ciertos casos sus aplicaciones generan ciertas distorsiones en los resultados, puntos que más adelante en el presente trabajo se precisarán. Todo esto es lo suficientemente factible para demostrar que se está presente ante un problema recurrente en todo trabajo de investigación y que la presencia de diversas metodologías que se han ido planteando a través de los años por diversos autores han propiciado la búsqueda de mejores métodos que puedan hacer frente a este problema, mejorando estimaciones y el tiempo con el que son obtenidas.

1.3 ANTECEDENTES NACIONALES E INTERNACIONALES

El afán de los investigadores por encontrar una solución que aminore las desventajas al poseer un conjunto de base de datos con información faltante, fomentaron las investigaciones en torno a este tema. Los trabajos de Wilks (1932), Scheuren (1976), Buck (1960), dieron origen a los métodos de imputación simple como el método de imputación por medias, imputación Hot Deck e imputación por regresión respectivamente. Los métodos fueron

ampliamente utilizados en diferentes artículos de investigación como los trabajos de Brown (1994), Enders & Bandalos (2001), Gleason & Staelin (1975) que utilizaron el método de imputación por medias, los de Ford (1983), Rao & Shao (1992), Kim & Fuller (2004), utilizando el método de imputación Hot Deck, y los artículos de Beale & Little (1975), Gleason & Staelin (1975), Kromrey & Hines (1994), Olinsky (2003), utilizando el método de imputación por regresión.

Otros autores plantearon el método de estimación por máxima verosimilitud utilizando el algoritmo EM, los orígenes del presente algoritmo datan en los trabajos realizados por Beale & Little (1975), Orchard & Woodbury (1972), Dempster et al. (1977), éste último cumplió un papel muy importante en el desarrollo del método el cual posteriormente fue utilizado en los trabajos de Jamshidian & Bentler (1999), Liang & Blenter (2004), McLachlan & Krishnan (1997). Finalmente el método de imputación múltiple fue desarrollado por Rubin (1987) y ha sido ampliamente utilizado en trabajos de Van Buuren (1999), Allison (2000), Royston (2004).

1.4 IMPORTANCIA DE LA INVESTIGACIÓN

La investigación en el presente trabajo refleja un aporte importante ya que presentará los métodos de estimación por Máxima Verosimilitud con el algoritmo EM y la Imputación Múltiple de datos para la estimación de parámetros, pasando por la metodología y conocimientos previos a la imputación. A la vez permitirá contrastar la exactitud de las estimaciones entre un conjunto de datos completo y conjunto de datos imputado por los

métodos planteados, finalmente se realizará una comparación de las estimaciones obtenidas de ambos métodos entre los casos de muestra completa y aquella con información faltante.

1.5 JUSTIFICACIÓN DE LA INVESTIGACIÓN

Es por ello que a través de este trabajo se pretende plantear a la estimación por Máxima Verosimilitud y a la Imputación Múltiple como posibles soluciones viables para hacer frente a la problemática de datos faltantes en un conjunto de datos, métodos que no presentan las desventajas existentes con respecto a los métodos de imputación tradicionales. Complementando dichos métodos con aplicaciones ejecutadas en el programa Matlab, buscando obtener resultados más precisos y una rápida ejecución en el caso de conjunto de datos grandes.

1.6 OBJETIVOS DE LA INVESTIGACIÓN

OBJETIVO GENERAL:

Comparación de estimaciones mediante simulación del algoritmo EM y la Imputación Múltiple para datos faltantes.

OBJETIVOS ESPECÍFICOS:

- Describir la metodología y procedimiento de la estimación por Máxima Verosimilitud mediante el algoritmo EM.
- Describir la metodología y procedimiento de la Imputación Múltiple para datos faltantes.

- Aplicar el algoritmo EM y la Imputación Múltiple a un conjunto de datos simulado con información faltante.
- Realizar una comparación de las estimaciones obtenidas con la muestra original simulada y con las obtenidas mediante el método de Listwise Deletion, el algoritmo EM y la Imputación Múltiple.
- Realizar una comparación de las estimaciones obtenidas con el algoritmo EM y la Imputación Múltiple.

1.7 DEFINICIONES IMPORTANTES

Concepto Bayesiano

La interpretación bayesiana de la probabilidad puede ser visto como una extensión de la lógica que permite el razonamiento con proposiciones cuya verdad o falsedad es incierto. Para evaluar la probabilidad de una hipótesis, la probabilística bayesiana especifica alguna probabilidad a priori, que se actualiza a la luz de nuevos y relevantes datos.

Concepto Frecuentista

La definición frecuentista consiste en definir la probabilidad como el límite cuando n tiende a infinito de la proporción o frecuencia relativa del suceso. Es imposible llegar a este límite, ya que no podemos repetir el experimento un número infinito de veces, pero si podemos repetirlo muchas veces y observar como las frecuencias relativas tienden a estabilizarse. Esta definición frecuentista de la probabilidad se llama también probabilidad a posteriori ya que sólo podemos dar la probabilidad de un suceso después de repetir y observar un gran número

de veces el experimento aleatorio correspondiente. Algunos autores las llaman probabilidades teóricas.

Data Completa

Conjunto de datos que contiene información completa sin presencia de valores faltantes, tiene un sentido hipotético debido a que en la práctica los casos de no respuesta son frecuentes. La data completa constará de dos componentes, la data observada y la data faltante denominadas $\mathbf{Y_{obs}}$, $\mathbf{Y_{fal}}$ respectivamente.

Distribución a priori

La distribución a priori cumple un papel importante en el análisis bayesiano ya que mide el grado de conocimiento inicial que se tiene de los parámetros en estudio, tanto así que si se tiene un conocimiento previo sobre los parámetros, este se traducirá en una distribución a priori.

Distribución a priori difusa o no informativa

Cuando nada es conocido sobre los parámetros, la selección de una distribución a priori adecuada será necesaria y más aun que ésta no influya sobre ninguno de los posibles valores de los parámetros en cuestión. Estas distribuciones a priori reciben el nombre de difusas o no informativas. En situaciones generales, para un parámetro θ el método más usado es el de Jeffreys.

Imputación

La imputación es la sustitución de valores no informados en una observación por otros. A veces es un paso necesario para poder tratar los datos con determinadas técnicas estadísticas de análisis. Idealmente, este análisis debería tener en cuenta el hecho de que algunos de los datos no son observados sino que han sido imputados.

Imputación múltiple

Consiste en asignar a cada valor faltante varios valores (m), generando m conjuntos de datos completos. En cada conjunto de datos completos se estiman los parámetros de interés y posteriormente se combinan los resultados obtenidos.

Imputación simple

Consiste en asignar un valor por cada valor faltante basándose en el valor de la propia variable o de otras variables, generando una base de datos completa.

Simulación

Una simulación por computadora o un modelo de simulación es un programa informático cuyo fin es crear un modelo abstracto de un determinado sistema. Las simulaciones por computadora se han convertido en una parte relevante y útil de los modelos matemáticos de muchos sistemas naturales de ciencias.

Datos faltantes

Son aquellos que no constan de información debido a cualquier acontecimiento, como por ejemplo errores en la transcripción de los datos o la ausencia de disposición a responder a ciertas cuestiones de una encuesta. Los datos pueden faltar de manera aleatoria o no aleatoria. Los datos faltantes aleatorios pueden perturbar el análisis de datos dado que disminuyen el tamaño de las muestras y en consecuencia la potencia de las pruebas de contraste de hipótesis mientras que los no aleatorios ocasionan una disminución de la representatividad de la muestra.

Método de Jeffreys

El método de Jeffreys (1961) sugiere que, si un investigador no tiene conocimiento del valor del parámetro θ, entonces a través de parametrizaciones de θ podrá obtenerse una expresión a cerca de θ dado los valores de Y, para obtener una priori invariante seguimos los siguientes pasos:

$$\Pr(\theta) \propto \sqrt{I(\theta)}$$

donde $I(\theta)$ es la matriz de información de Fisher

$$I(\theta) = -E_\theta\left[\frac{\partial^2 Lnf(y|\theta)}{\partial\theta^2}\right]$$

Si $\theta = (\theta_1, \theta_2, \ldots, \theta_n)'$ es un vector, entonces:

$$\Pr(\theta) \propto \sqrt{\det I(\theta)}$$

donde $I(\theta)$ es la matriz de información de Fisher de orden pxp.

El elemento (i j) de esta matriz es:

$$I_{ij} = -E_0\left[\frac{\partial^2 Lnf(y|\theta)}{\partial\theta_i\,\partial\theta_j}\right]$$

CAPÍTULO II IMPUTACIÓN DE DATOS FALTANTES

2.1 HISTORIA DE LOS MÉTODOS DE IMPUTACIÓN DE DATOS FALTANTES

El problema de trabajar con un conjunto de datos con información faltante inicialmente fue solucionado con métodos de fácil aplicación entre los que se encontraba el Listwise Deletion (método que ignora aquellas unidades informativas con ausencia de datos), debido a las desventajas presentadas en cuanto a pérdidas grandes de información en casos de conjunto de datos de gran volumen es que surgieron nuevos métodos que no presentaran los mismos inconvenientes, naciendo así los métodos de imputación de datos.

Los nuevos aportes en la imputación de datos se realizaron en el año 1932 por Wilks, quien utilizó la sustitución de datos faltantes apoyándose en la media de las variables presentes, posteriormente y ante los reportes de introducción de sesgos por parte de esta metodología, es que se originó un interés por parte de los investigadores de la década de los setenta; quienes en la búsqueda de correcciones y de nuevos métodos plantearon la imputación por regresión (Buck 1960), cuya idea básica era el reemplazo o sustitución de los datos faltantes por valores predictivos de una ecuación de regresión; dicho método también presentó inconvenientes con respecto a la aparición de sesgos en las varianzas y covarianzas.

El método de Hot Deck, surgió en los censos Bureau usadas en sus datas de uso público, definiéndolas como un conjunto de técnicas que imputan datos faltantes con valores de una similar unidad informativa, muy a pesar de haberse obtenido mejores resultados en las

aplicaciones prácticas este método resultó no ser el adecuado para la estimación de medidas de asociación debido a la introducción de sesgo en las correlaciones y coeficientes estimados.

Ya con el avance tecnológico de aquellos años se utilizaron los sistemas computacionales en la realización de investigaciones sobre estos temas, entre los principales autores de los últimos años que han hecho grandes investigaciones referentes a la imputación figuran Kalton, Kasprzyk, Little, Rubin. Éste último en 1976 propuso un marco conceptual para el análisis de datos faltantes sustentado en métodos de inferencia estadística. Posteriormente, la aparición del algoritmo Expectation Maximization (EM) permitió generar estimadores robustos a partir de la aplicación de la estimación por máxima verosimilitud EMV (Dempster, Laird, y Rubin, 1977), en donde las observaciones faltantes se asumen como variables aleatorias y los datos imputados se generan.

2.2 MECANISMO DE DATOS FALTANTES

Rubin (1976), advierte que la ausencia de datos debe analizarse como un fenómeno estocástico y que éste juega un rol importante en cuanto al uso de metodologías. Rubin consideró tres mecanismos de data faltante que describen como la probabilidad de los valores faltantes se relaciona con los datos si es que esto ocurre, esta tipología propuesta por Rubin es ampliamente utilizada hasta la actualidad. Con el fin de comprender los mecanismos de datos faltantes es necesario denotar las siguientes expresiones.

Sea: $Y = (Y_{obs}, Y_{fal})$, la variable aleatoria con distribución de probabilidad conjunta

Y_{obs}, el vector de datos observados y Y_{fal}, el vector de datos faltantes

Los mecanismos de datos faltantes definidos por Rubin son los siguientes:

- **MAR (Missing at Random) - Perdidos al Azar.**

  Los datos perdidos al azar se generan cuando la probabilidad de pérdida de una variable faltante (Y_{fal}), está relacionada con una o más variables observables (Y_{obs}) y no con la variable faltante misma (Y_{fal}). Es decir la ausencia de datos está asociada a variables observables presente en el conjunto de datos.

- **MCAR (Missing Completely at Random) - Perdidos Completamente al Azar.**

  Los datos perdidos completamente al azar se generan cuando la probabilidad de pérdida de una variable faltante (Y_{fal}) no está relacionada con las variables observables (Y_{obs}) ni con las variables faltantes (Y_{fal}). Es decir la ausencia de la información no ha sido originada por ninguna variable presente en el conjunto de datos.

- **NMAR (Not Missing at Random) – No Perdidos al Azar.**

  Los datos no perdidos al azar se producen cuando la probabilidad de pérdida de una variable faltante (Y_{fal}) está sólo relacionado con la misma variable faltante (Y_{fal}).

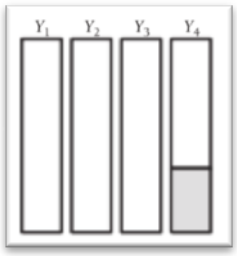
2.3 PATRONES DE DATOS FALTANTES

Los distintos procedimientos de imputación requieren de ciertos supuestos acerca del patrón de datos faltantes. Si el conjunto de datos se interpreta como una matriz, en donde las filas son las unidades informativas y las columnas representan a las variables de interés, la elección del método de imputación debiera tener en cuenta el comportamiento de los datos

omitidos, ya que el análisis visual permite identificar patrones como los que se muestran a continuación.

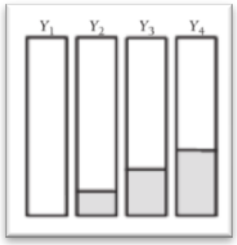
- **PATRÓN UNIVARIADO:**

  Cuando los datos faltantes se concentran en una variable, tal es el caso del gráfico donde la ausencia de información se presenta en la variable Y_4.



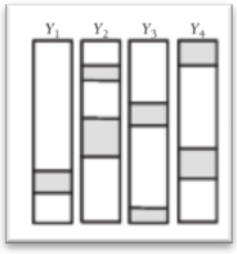
- **PATRÓN MONÓTONO:**

  Cuando los datos faltantes poseen un patrón escalonado, este comportamiento de ausencia de información es característico de los estudios longitudinales.



- **PATRÓN ALEATORIO:**

  Cuando en cualquier celda pueden existir datos faltantes es decir, las omisiones de información no están dispuestas en forma predeterminada.



2.4 SUPUESTOS

Como cualquier método estadístico los métodos de imputación en específico el múltiple, requieren del cumplimiento de ciertos supuestos fundamentados en su metodología (Rubin 1987), supuestos que en la práctica y en la mayoría de los casos no son fácilmente de

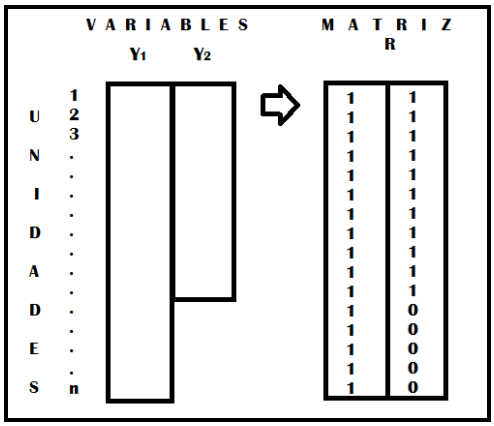
demostrar y cumplir. Para tener un básico entendimiento de los mismos se mencionan a continuación:

- **EL MODELO DE DATA COMPLETA**

  Para generar imputaciones de los valores perdidos, se debe definir un modelo de probabilidad de la data completa $Y = (Y_{obs}, Y_{fal})$ que considera a los valores observados y faltantes. Por lo tanto se considerara un conjunto de datos rectangular cuyas filas pueden ser modeladas como independientes e idénticamente distribuidas (iid) desde algunas distribuciones de probabilidad multivariada. Sea $\mathbf{Y}$ la matriz $n \times p$ de datos completa, que presenta variables observables como faltantes y consideremos a $\mathbf{y_i}$ como la fila i de la variable $Y = (y_1, y_2, \dots, y_n)^T$, donde $y_i = (y_{i1}, y_{i2}, \dots, y_{ip})^T$ es una muestra aleatoria desde una distribución de probabilidad p-dimensional multivariada $P(Y|\theta)$, por tanto las filas son representadas por $y_i (i = 1,2, \dots, n)$ y las columnas de Y como variables denotadas por $Y_j (j = 1,2, \dots, p)$. Además se define una matriz indicador de pérdida $n \times p$, siendo $R = (r_{ij})$ la distribución de dicha matriz,

$$r_{ij} \begin{cases} 1 & si\ y_{ij}\ es\ observable\ (Y_{obs}) \\ 0 & si\ y_{ij}\ es\ faltante\ (Y_{fal}) \end{cases}$$

  Por ejemplo si planteamos un conjunto de datos bivariado con sólo una variable con valores faltantes se presentará un patrón univariado con matriz indicador correspondiente al siguiente gráfico.



Por el supuesto anteriormente mencionado, la distribución de probabilidad conjunta de las variables respuesta y de las variables indicadoras de pérdida se puede expresar como:

$$P(Y, R|\theta, \xi) = P(Y|\theta)\, P(R|\xi, Y)$$

Donde $P(Y|\theta)$ es la distribución marginal de las variables respuestas y $P(R|\xi, Y)$ es la distribución condicional de pérdida con respecto a las variables respuesta.

- **IGNORABILIDAD**

  La ignorabilidad requiere asumir el cumplimiento de dos supuestos, el primero hace referencia a que el mecanismo de datos faltantes sea considerado perdido al azar (mecanismo **MAR**) y el segundo a la distinción de los parámetros que asume que los parámetros θ del modelo de datos y ξ del mecanismo de datos faltantes son diferentes, desde una perspectiva frecuentista la distinción de parámetros asume que el espacio de parámetros comunes de (θ, ξ) es considerado como el producto cartesiano entre los espacios individuales de los respectivos parámetros, mientras que desde la perspectiva

bayesiana atribuye que para cualquier distribución conjunta a priori aplicada a (θ, ξ) puede ser factorizada en sus marginales a priori independientes respectivamente.

Cuando ambos supuestos se cumplen se asume que el mecanismo de datos faltantes puede ser ignorable (Little y Rubin 1987; Rubin 1987), en este mismo sentido el mecanismo de pérdida también puede ser ignorado cuando realicemos inferencias basado en probabilidad o en estadística bayesiana.

- **LA DISTRIBUCIÓN DE PROBABILIDAD DE LOS DATOS OBSERVADOS $\mathbf{L(\theta|\, Y_{obs})}$**

Bajo el cumplimiento del supuesto de ignorabilidad donde el mecanismo de datos faltantes es ignorable, no necesitaremos considerar el modelo *R* ni el parámetro ξ cuando realicemos inferencias de θ, por tanto la distribución de probabilidad de los datos observados se describirá como:

$$P(Y_{obs}, R|\theta, \xi) = \int P(Y_{obs}, Y_{fal}, R|\theta, \xi)\, dY_{fal}$$

Donde por el supuesto de distinción de parámetros se puede expresar a $P(Y_{obs}, R|\theta, \xi)$ como:

$$P(Y_{obs}, R|\theta, \xi) = \int P(Y_{obs}, Y_{fal}|\theta)\, P(R|Y_{obs}, Y_{fal}, \xi)\, dY_{fal}$$

y según el mecanismo MCAR se obtiene

$$P(Y_{obs}, R|\theta, \xi) = \int P(Y_{obs}, Y_{fal}|\theta)\, P(R|\xi)\, dY_{fal}$$

$$P(Y_{obs}, R|\theta, \xi) = P(R|\xi) \int P(Y_{obs}, Y_{fal}|\theta)\, dY_{fal}$$

$$P(Y_{obs}, R|\theta, \xi) = P(R|\xi) P(Y_{obs}|\theta)$$

Según el mecanismo MAR se obtiene

$$P(Y_{obs}, R|\theta, \xi) = \int P(Y_{obs}, Y_{fal}|\theta)\, P(R|Y_{obs}, \xi) dY_{fal}$$

$$P(Y_{obs}, R|\theta, \xi) = P(R|Y_{obs}, \xi) \int P(Y_{obs}, Y_{fal}|\theta)\, dY_{fal}$$

$$P(Y_{obs}, R|\theta, \xi) = P(R|Y_{obs}, \xi) P(Y_{obs}|\theta)$$

Por lo tanto,

$P(Y_{obs}, R|\theta, \xi) = P(R|\xi) P(Y_{obs}|\theta)$ Según mecanismo MCAR

$P(Y_{obs}, R|\theta, \xi) = P(R|Y_{obs}, \xi)\ P(Y_{obs}|\theta)$ Según mecanismo MAR

Asumiendo de acuerdo a la metodología los supuestos de ignorabilidad del mecanismo de pérdida, es que se obtiene una distribución de probabilidad de los datos observados $P(Y_{obs}, R|\theta, \xi)$ integrada por los factores; $P(R|\xi)$ y $P(Y_{obs}|\theta)$ en el caso MCAR y respectivamente en el caso MAR a $P(Y_{obs}|\theta)$ y $P(R|Y_{obs}, \xi)$ que hacen referencia a los parámetros de interés θ y ξ, asumiendo que las inferencias basadas en θ no se verán afectadas por ξ por $P(R|\xi)$ ni por $P(R|Y_{obs}, \xi)$. por tanto es que en algunos métodos como la estimación por máxima verosimilitud puede ser realizado sin considerar el mecanismo de datos faltante.

De esta manera, la distribución de probabilidad de los datos observados puede ser reemplazado por su distribución marginal para propósitos de inferencias de θ y desde el enfoque frecuentista e ignorando el mecanismo de pérdida, la distribución de probabilidad de datos observados es proporcional a:

$$L(\theta|\, Y_{obs}\,) \;\propto\; P(Y_{obs}|\theta)$$

- **LA DISTRIBUCIÓN POSTERIOR DE LA DATA OBSERVADA $\mathbf{P(\theta \mid Yobs)}$**

Desde la perspectiva bayesiana, todas las inferencias están fundamentadas en la distribución de probabilidad a posteriori y su obtención se basa en el teorema de Bayes:

$$P(\theta|Y) = c\, P(Y|\theta)P(\theta)$$

Donde:

$P(\theta)$ representa la distribución a priori de θ

$P(\theta|Y)$ representa la distribución a posteriori de θ dado Y

c representa una constante normalizada necesaria para que $P(Y|\theta)$ sume o integre uno

Por tanto la distribución posterior bajo el teorema de Bayes puede determinarse como,

$$P(\theta\,,\xi|Y_{obs}, R\,) = \; k^{-1}\, P(Y_{obs}\,, R\,|\theta\,,\xi\,)\pi\,(\theta\,,\xi\,)$$

Donde $\pi\,(\cdot)$ denota la distribucion a prior aplicada a $(\theta\,,\xi)$ y k una constante de normalización.

Bajo el supuesto MCAR y MAR, podríamos expresar a la distribución a posteriori como

$P(\theta\,,\xi|Y_{obs}, R\,) = \; k^{-1}\, P(R|Y_{obs}\,,\xi\,)\;\; P(Y_{obs}|\theta)\;\pi\,(\theta\,,\xi\,)$ MCAR

$P(\theta\,,\xi|Y_{obs}, R\,) = \; k^{-1}\, P(R|\xi\,)\;\; P(Y_{obs}|\theta)\;\pi\,(\theta\,,\xi\,)$ MAR

Tenemos que las inferencias bayesianas de θ están basadas en la distribución marginal posterior que se obtendrán integrando esta función sobre el parámetro ξ, y bajo el supuesto de distinción es que se obtendrán los factores de la distribución prior

$$\pi(\theta, \xi) = \pi_\theta(\theta)\pi_\xi(\xi)$$

$$P(\theta, \xi | Y_{obs}, R) = k^{-1} P(R|\xi)\ P(Y_{obs}|\theta)\ \pi_\theta(\theta)\pi_\xi(\xi) \qquad \text{MCAR}$$

$$P(\theta, \xi | Y_{obs}, R) = k^{-1} P(R|Y_{obs}, \xi)\ P(Y_{obs}|\theta)\ \pi_\theta(\theta)\pi_\xi(\xi) \qquad \text{MAR}$$

De ahí la distribución marginal posterior de θ,

$$P(\theta \mid Yobs, R) = \int P(\theta, \xi \mid Yobs, R)\, d\xi$$

$$P(\theta \mid Yobs, R) = P(Y_{obs}|\theta)\ \pi_\theta(\theta) \int P(R|\xi)\ \pi_\xi(\xi)\, d\xi \qquad \text{MCAR}$$

$$P(\theta \mid Yobs, R) = P(Y_{obs}|\theta)\ \pi_\theta(\theta) \int P(R|Y_{obs}, \xi)\ \pi_\xi(\xi)\, d\xi \qquad \text{MAR}$$

Donde se obtiene para ambos casos

$$P(\theta \mid Yobs, R) = P(Y_{obs}|\theta)\ \pi_\theta(\theta)$$

donde la proporcionalidad depende de un factor multiplicativo que no implica θ. Por tanto, $P(\theta \mid Yobs, R) = P(\theta \mid Yobs)$, además bajo ignorabilidad toda la información sobre θ se resume en la distribución posterior, que ignora la falta de datos de mecanismo.

$$P(\theta \mid Yobs) \propto L(\theta \mid Yobs)\pi_\theta(\theta)$$

- **LA DISTRIBUCIÓN A PRIORI $\pi(\theta)$**

  En inferencia estadística Bayesiana, la distribución de probabilidad a priori de un parámetro es la distribución que describe todo conocimiento subjetivo acerca de la probabilidad relativa de diferentes valores de parámetros antes de la toma de información, en otras palabras la distribución a priori es todo conocimiento previo o información relevante que se tenga sobre los parámetros, siendo así posible plantear diferentes distribuciones a priori por cada información previa que se tenga lo que nos hará obtener diferentes resultados en la distribución a posteriori, diferencias que dejarán de ser importantes cuando el tamaño de muestra sea moderadamente grande.

  Para los casos en que no se tenga conocimientos previos de los parámetros, la distribución a priori pasa a ser considerada como no informativa, y su elección tendrá un sentido específico ya que se deberá escoger una distribución que no distorsione los valores de los parámetros, situación en la cual el método más común a utilizar es el método de Jeffreys, método por el cual se obtiene una distribución a priori como resultado de una parametrización de θ obtención que queda explicado en las definiciones importantes del capítulo I de la presente tesis.

2.5 MÉTODOS DE IMPUTACIÓN DE DATOS CONVENCIONALES

A continuación los métodos de imputación más utilizados a lo largo del tiempo:

- **IMPUTACIÓN POR LA MEDIA**

  Este método, propuesto por primera vez por Wilks (1932), es posiblemente uno de los procedimientos de imputación más antiguo y más sencillos. Los valores faltantes de una variable se sustituyen por la media obtenida de las unidades observadas de la variable en mención y su aplicación asume que los datos faltantes siguen un mecanismo MCAR. Este método ocasiona que se preserve el valor medio (la media) afectando la distribución de probabilidad de la variable imputada, otra desventaja que ocasiona es que se atenúan las medidas de asociación como la correlación entre variables generando una distorsión en la relación de las mismas a la vez que afectan a la covarianza. Para Acock (2005), este es el peor de los procedimientos de imputación, y por tanto no recomienda su uso.

- **IMPUTACIÓN HOT DECK**

  El procedimiento Hot Deck es una metodología que consiste en el reemplazo de aquellos valores faltantes por valores de similares entrevistados, es decir que compartan las mismas características. De esta metodología surgieron derivados como Hot Deck Secuencial, Hot Deck vecino más cercano, etc. Si bien es cierto fue otro novedoso método, no evitó los problemas presentados en cuanto a sesgos en las correlaciones estimadas, en los coeficientes de regresión estimadas y en los errores estándares.

- **IMPUTACIÓN POR REGRESIÓN**

  Es un método propuesto por primera vez por Buck (1960), la idea básica de esta metodología es utilizar la información de las variables con información completa para reemplazar las faltantes, empleando modelos de regresión para imputar la información.

Los valores faltantes se sustituyen por valores predictivos obtenidos a partir de una regresión entre un conjunto de variables predictoras X y un conjunto de variables explicativas Y.

Si bien es cierto que esta metodología resultó ser mejor que la imputación por medias, ocasiona resultados inversos a ésta, mientras que el anterior método atenúa las correlaciones entre variables; el método de imputación por regresión obtiene correlaciones altas casi perfecta en el caso de correlacionar los valores predictivos y el subconjunto imputado pronosticado, es decir sobrestima las correlaciones aún cuando la data es considerada MCAR.

Otra de las desventajas que posee este método es que los valores imputados se ajustarán directamente en la recta de regresión produciendo con esto poca variabilidad y por tanto atenuando la varianza o covarianza.

- **IMPUTACIÓN POR REGRESIÓN ESTOCÁSTICA**

  El método de imputación por regresión estocástica sugiere el mismo proceso de imputación por regresión pero con la adición a la ecuación de regresión de un valor aleatorio con distribución normal con media cero y varianza igual a la varianza residual de la regresión entre variables. Esta incorporación añade la variabilidad que se pierde en el método de imputación por regresión ocasionando la obtención de parámetros estimados insesgados.

2.6 CONSIDERACIONES PARA LA IMPUTACIÓN DE DATOS FALTANTES

Es importante recalcar que si bien es cierto no existe criterios fijos previos para analizar un conjunto de datos que tiene presencia de información faltante, la marcada experiencia de algunos autores precisan ciertas consideraciones previas que deben ser tomadas en cuenta previo a la imputación de datos faltantes.

Según Goicoechea (2002), las consideraciones que se deben tener en cuenta para seleccionar el método de imputación son las siguientes:

- **IMPORTANCIA DE VARIABLE A IMPUTAR**

  Dándose el caso que la variable a imputar sea de suma importancia es que se debe tener en cuenta el tipo de modelo de imputación a elegir, debido a que si la variable imputada está considerada parte de nuestros objetivos de investigación, es que se debe tener sumo cuidado en la elección.

- **TIPO DE VARIABLE A IMPUTAR**

  Es importante definir si la o las variables son continuas o categóricas, nominal u ordinal, ya que esta característica define en parte el método de imputación a usar.

- **PARÁMETROS QUE SE DESEAN ESTIMAR**

  En el caso en que solamente nuestro objetivo sea obtener promedios, se pueden aplicar los métodos más sencillos como el de la imputación por medias, sin embargo si quisiéramos obtener también las estimaciones de las varianzas o covarianzas este método

nos ocasionará problemas ya que subestimará sus valores reales, en estos casos es necesario métodos más elaborados, dificultándose la situación cuando hay una elevada tasa de no respuesta.

- **TASA DE NO RESPUESTA**

  La tasa de no respuesta es otra de las consideraciones a tomar en cuenta al escoger un método de imputación, si ésta es mínima no se ve necesario aplicar método imputables; pero cuando el número de ítems perdidos sea considerable si podría generar más de un problema como la pérdida de representatividad de la muestra, la distorsión de frecuencias, los sesgos en la estimación y aumento de error de muestreo. Según Laaksonen (2000) considera la tasa de no-respuesta elevada cuando supera un tercio del total del tamaño muestral y en la práctica otros muchos investigadores hablan de pérdidas máximas entre 1 y 20 por ciento, lo cierto es que en la práctica dependerá mucho de la precisión del estudio, el área y objetivos de la investigación para considerar utilizar técnicas de imputación.

- **INFORMACIÓN AUXILIAR DISPONIBLE**

  La imputación puede mejorar al emplear información auxiliar disponible, ya que con ella el deducir y el reemplazar los valores ausentes con características similares al grupo o unidad informante puede ayudar a la exactitud de las estimaciones.

También otras de las consideraciones a tomar en cuenta son el mecanismo de datos faltantes (Rubin 1987) y el patrón de ausencia de datos, ya que de estos dos criterios dependerá

también el método de imputación a utilizar. Si bien es cierto, se considera un método de imputación de acuerdo a los criterios antes mencionados, es importante seleccionar otros métodos de imputación para poder contrastar los resultados que se obtienen y de esta manera asegurar nuestra elección.

CAPÍTULO III ESTIMACIÓN POR MÁXIMA VEROSIMILITUD

3.1 ESTIMACIÓN POR MÁXIMA VEROSIMILITUD

Muchos de los procedimientos estadísticos suponen que los datos siguen algún tipo de modelo matemático y hay algunos en los que se desconoce sus parámetros. Con el objetivo de obtener valores estimados de los parámetros desconocidos es que se aplica diferentes métodos de estimación de entre los cuales probablemente el más versátil debido a su aplicación en gran cantidad de situaciones; es el método de máxima verosimilitud.

El uso del método de estimación por máxima verosimilitud data de los años 50's en Anderson 1957, Edgett 1956, Hartley 1958, Lord 1955, con pocas aplicaciones prácticas debido al alcance limitado de la época.

El procedimiento de la estimación por máxima verosimilitud conocida como EMV, supone que los datos completos siguen un determinado modelo multivariante; con el cual se podrá realizar estimaciones verosímiles de los parámetros, pero la pregunta que nos hacemos es que sucede cuando tenemos datos faltantes y es allí cuando el método de máxima verosimilitud busca identificar los parámetros estimados más probables de haber producido el conjunto de datos. Conceptualmente el proceso de estimación es el mismo con la presencia de datos faltantes o sin ella; sin embargo la presencia de datos faltantes puede complicar el uso de la función de verosimilitud de los datos faltantes y a los ajustes que se requieran por los cuales se generan procesos o algoritmos iterativos.

3.2 PROCEDIMIENTO DE LA ESTIMACIÓN POR MÁXIMA VEROSIMILITUD

El procedimiento general para estimar los parámetros de un modelo con valores faltantes haciendo uso del método EMV se inicia estimando los parámetros del modelo con los datos completos con la función de máxima verosimilitud, estos parámetros posteriormente se utilizan para predecir los valores faltantes que se presenten, una vez que se obtenga un conjunto de datos completos ya con los valores reemplazados nuevamente se obtiene nuevos parámetros maximizando la verosimilitud de la muestra completa, iniciándose un proceso iterativo, el proceso será repetitivo hasta que se logre llegar a la convergencia, y con esta la obtención de la máxima probabilidad.

Para estimar los parámetros de una distribución normal multivariante con presencia de datos faltantes utilizamos la función de densidad conjunta de las observaciones denotada:

Sea $y_1, y_2, \dots, y_n$ una muestra aleatoria simple donde $y_i \sim N_p(\mu, \Sigma)$ para obtener los estimadores de máxima verosimilitud de μ, Σ se define la función de densidad conjunta de las observaciones

$$f(Y \mid \mu, \mathrm{V}) = \prod_{i=1}^{n} |\mathrm{V}|^{-\frac{1}{2}} \ (2\pi)^{-p/2} \ \exp\left\{-\frac{1}{2}\ (y-\mu)'\ \mathrm{V}^{-1}(y-\mu)\right\}$$

y la función de verosimilitud será despreciando las constantes

$$l(\mu, \mathrm{V} \mid Y) = -\frac{n}{2}\, log|\mathrm{V}| - \frac{1}{2}\sum_{i=1}^{n}(y-\mu)'\,\mathrm{V}^{-1}\,(y-\mu)$$

Observemos que la función de verosimilitud así escrita es siempre negativa, ya que tanto el determinante como la forma cuadrática son positivos por ser definida positiva la matriz V.

Esta función nos indica el apoyo o soporte que reciben los posibles valores de los parámetros dados los valores muestrales observados. Cuanto mayor sea esta función (menos negativa) para unos valores de los parámetros, mayor será la concordancia entre estos parámetros y los datos. Llamando $\bar{y} = \sum_{i=1}^{n} y_i/n$ al vector de medias muestrales y escribiendo $(y_i - \mu) = (y_i - \bar{y} + \bar{y} - \mu)$ y desarrollando la forma cuadrática

$$\sum_{i=1}^{n} (y-\mu)' V^{-1} (y-\mu) = \sum_{i=1}^{n} (y-\bar{y})' V^{-1} (y-\bar{y}) + n\,(\bar{y}-\mu)' V^{-1} (\bar{y}-\mu)$$

ya que $\sum_{i=1}^{n} (y-\bar{y}) = 0$, desarrollando el primer término de esta expresión tenemos que

$$tr\left(\sum_{i=1}^{n} (y-\bar{y})' V^{-1} (y-\bar{y})\right) = \sum_{i=1}^{n} tr\,[(y-\bar{y})' V^{-1} (y-\bar{y})]$$

$$= \sum_{i=1}^{n} tr\,[V^{-1}\ (y-\bar{y})\ (y-\bar{y})'] = tr(\ V^{-1}\ \sum_{i=1}^{n} (y-\bar{y})(y-\bar{y})'\ )$$

reemplazando los valores correspondientes obtenemos a expresión

$$l(\mu, V \mid Y) = -\frac{n}{2} log|V| - \frac{1}{2} tr\left(V^{-1} \sum_{i=1}^{n} y_i\, y_i'\right) - \frac{n}{2}\, \mu' V^{-1} \mu + \mu' V^{-1} \sum_{i=1}^{n} y_i$$

en la cual la estimación MV cuando se tiene la muestra completa es

$$\hat{\mu} = \sum_{i=1}^{n} \frac{y_i}{n} \quad y \quad \hat{V} = \frac{\sum_{i=1}^{n} y_i\, y_i'}{n} - \hat{\mu}\hat{\mu}'$$

Ahora considerando la función de densidad conjunta en términos de la data completa Y se define,

$$p(Y|\theta) = p\big(Y_{obs}, Y_{fal}\,\big|\theta\big), \qquad \theta \in \Theta$$

$$p(Y|\theta) = p\left(Y_{fal}\middle|Y_{obs}, \theta\right) p(Y_{obs}|\theta) \;, \qquad \theta\epsilon\,\Theta$$

La función de máxima verosimilitud está dado por,

$$l(Y \mid \theta) = log\, p\left(Y_{fal}\middle|Y_{obs}, \theta\right) + l(\theta|Y_{obs})$$

Donde:

$l(Y \mid \theta)$: Log – verosimilitud de la data completa.

$l(\theta|Y_{obs})$: Log – verosimilitud de la data observada.

$log\, p(Y_{fal}|Y_{obs}, \theta)$: Distribución predictiva de la data faltante dado θ.

Debido a los difíciles cálculos y procedimientos del método de estimación por máxima verosimilitud en los casos de manejo de grandes conjuntos de información, en el que por el año 1977 se desarrollaron modernas técnicas entre las que el algoritmo iterativo EM (Dempster, Laird y Rubin) logró ejecutarse con eficiencia.

Para plantear este algoritmo es necesario tomar en cuenta las siguientes definiciones de términos:

$Y_{obs} \in \mathbb{R}^n$: Vector n – dimensional de cantidades observadas.

$Y_{fal} \in \mathbb{R}^m$: Vector m – dimensional de cantidades no observadas.

$Y \in \mathbb{R}^{n+m}$: Data completa.

$Y = (Y_{obs}, Y_{fal})$

3.3 ALGORITMO EM

El algoritmo EM es una técnica general para ajustar modelos de información faltante, procedimiento que se basa en la relación entre la información faltante y los parámetros desconocidos de un modelo de distribución. Para dar una mejor explicación del tema Schafer (1999) mencionó que si se conocieran los valores perdidos la estimación de los parámetros del modelo sería inmediato y de la misma manera si se conocieran los parámetros del modelo podrían obtenerse las predicciones insesgados de los valores faltantes, de esto trata el algoritmo EM, de un método iterativo que toma la interdependencia entre estos dos procesos.

El método se inicia obteniendo las predicciones de los valores faltantes basados en los valores iniciales que establece el investigador o analista, estas predicciones nuevamente son utilizadas para la obtención de un nuevo parámetro estimado repitiéndose este proceso reiteradas veces hasta que esta secuencia converge a un punto máximo.

3.4 FORMA GENERAL DEL ALGORITMO EM

La forma general del algoritmo EM repite los siguientes pasos hasta converger:

1. Fijar i = 0 e inicializar θ con un $\theta^{(0)}$ arbitrario.

   Partiendo de un estimador inicial $\hat{\theta}^{(i)}$, en la primera iteración $(i = 0)$.

2. **PASO E:** Calcular $Q(\theta|\theta^{(i)})$ utilizando $\theta^{(0)}$ y los datos observados de Y_{obs} para estimar la distribución de Y_{fal}.

   Se calcula a través de la esperanza matemática de las funciones de los valores faltantes que aparecen en la función de verosimilitud completa, $l(\theta\,|Y_{obs}\,,\,Y_{fal}\,)$ con respecto a la distribución Y_{fal} dados los valores $\hat{\theta}^{(i)}$ y los datos observados Y_{obs} .

$$l^{*}(\theta|Y_{obs}) = E_{Y_{fal}|\hat{\theta}^{(i)}}(l(\theta\,|Y_{obs}\,,Y_{fal}\,))$$

   el resultado obtenido de esta operación se denomina el paso E (Expectation o Predicción) del algoritmo.

   Para los casos en que $l(\theta\,|Y_{obs}\,,\,Y_{fal}\,)$ sea una función lineal de Y_{fal} este procedimiento nos conducirá a reemplazar en esta función a los Y_{fal} por las esperanzas dados los parámetros.

   Según el artículo de Dempster, Laird y Rubín, para maximizar la verosimilitud bastará encontrar un conjunto de parámetros $\theta^{(i+1)}$ que maximicen $Q(\theta|\theta^{(i)})$, es decir:

$$Q(\theta^{(i+1)}|\theta^{(i)}) \geq Q(\theta^{(i)}|\theta^{(i)})$$

donde $Q\left(\theta\middle|\theta^{(i)}\right) = E_{Y_{fal}|Y_{obs},\ \hat{\theta}^{(i)}}(l(\theta\ |Y_{obs}\ , Y_{fal}\ ))$

Entonces también quedará minimizado:

$$H\left(\theta^{(i)}\middle|\theta^{(i)}\right) \geq H\left(\theta^{(i+1)}\middle|\theta^{(i)}\right)$$

donde

$$H\left(\theta\middle|\theta^{(i)}\right) = E_{Y_{fal}|Y_{obs},\ \hat{\theta}^{(i)}}(log\ p\ (\ Y_{fal}|Y_{obs}\ , \theta\ ))$$

Y por lo tanto se tendrá:

$$l(\theta^{(i+1)}\ |\ Y_{obs}) \geq l(\theta^{(i)}\ |\ Y_{obs})$$

Como $H\left(\theta^{(i)}\middle|\theta^{(i)}\right)$ representa la función logaritmo de verosimilitud del conjunto total $l(\theta\ |\ Y)$, entonces:

$$l(\theta^{(i+1)}\ |\ Y) \geq l(\theta^{(i)}\ |\ Y) \rightarrow l(\theta^{(i+1)}\ |\ Y_{obs}) \geq l(\theta^{(i)}\ |\ Y_{obs})$$

Es decir si se maximiza el logaritmo de la verosimilitud del conjunto total de datos, se estará maximizando el logaritmo de la verosimilitud del conjunto observado de datos.

3. **PASO M:** Encontrar $\hat{\theta}^{(i+1)}$ del espacio **Θ** tal que $Q\left(\theta\middle|\theta^{(i)}\right)$ quede maximizado.

   Posteriormente se procederá a maximizar la función obtenida $l^{*}(\theta|Y_{obs})$ con respecto a θ, conocido como el paso M (Maximization o maximización) del algoritmo. Este paso M equivale a maximizar la verosimilitud completa donde ya se han reemplazado los valores faltantes por estimadores.

4. Si $\left\|\hat{\theta}^{(i+1)} - \hat{\theta}^{(i)}\right\| \neq 0$ incrementa i y volver a paso 2.

   Con el valor obtenido en el paso M $\hat{\theta}^{(i+1)}$, se procede a ejecutar el paso E y se itera hasta lograr la convergencia, es decir hasta que se haya obtenido una diferencia lo suficientemente pequeña entre $\left\|\hat{\theta}^{(i+1)} - \hat{\theta}^{(i)}\right\|$.

3.5 ALGORITMO EM APLICADO A POBLACIONES NORMALES CON DATOS FALTANTES.

El método de estimación de máxima verosimilitud aplicado en los casos de datos faltantes utilizando el algoritmo EM para poblaciones normalmente distribuidas es explicado en el punto precedente:

Comenzaremos calculando un estimador inicial con los datos disponibles. Sean $\hat{\mu}^{(0)}$ y $\hat{V}^{(0)}$ estos estimadores iniciales. Tomamos $\hat{\mu}^{(i)} = \hat{\mu}^{(0)}$ e $\hat{V}^{(i)} = \hat{V}^{(0)}$ e iteramos entre los pasos:

Paso E: Se calcula la esperanza de la función de verosimilitud completa definida respecto a los valores faltantes Y_{fal} dados los parámetros $\hat{\theta}^{(i)} = (\hat{\mu}^{(i)}, \hat{V}^{(i)})$ e Y_{obs}. En esta función los datos faltantes aparecen en dos términos. El primero es $\mu' V^{-1} \sum_{i=1}^{n} y_i$ donde aparecen en forma lineal por lo que tendremos simplemente que sustituir los datos faltantes por sus estimaciones. El segundo es $tr(V^{-1} \sum_{i=1}^{n} y_i \, y_i')$ donde tendremos que sustituir las expresiones $y_i \, y_i'$ por las estimaciones respectivas. Comencemos con el primer término, tomamos esperanzas de los parámetros y los datos conocidos, esto implica sustituir y_i para $i > m$ por $E\left(y_i \middle| Y_{obs}, \hat{\theta}^{(i)}\right)$.

Para visualizar el proceso, planteamos una distribución normal bivariada, donde $y_i = [y'_{1i}, y'_{2i}]'$ la cual se observa de manera parcial, es decir y'_{1i} es una variable completamente observada (Y_{obs}) mientras que y'_{2i} es parcialmente observada (Y_{fal}). Entonces al obtener la esperanza del primer término en mención $E(y_i|Y_{obs}, \hat{\theta}^{(i)})$, la cual depende de los valores observados de y_{1i} y será igual a la esperanza condicionada $E(y_{2i}|Y_{1i}, \hat{\theta}^{(i)})$.

Esta esperanza se calcula, por regresión mediante:

$$E(y_i|Y, \hat{\theta}^{(i)}) = E(y_{2i}|y_{1i}, \hat{\theta}^{(i)}) = \hat{y}_{2i.1}^{(i)} = \hat{\mu}_2 + \hat{V}_{12}^{(i)} (\hat{V}_{11}^{(i)})^{-1} (y_{1i} - \hat{\mu}_1^{(i)})$$

donde las expresiones pertenecen al vector de medias y la matriz de covarianzas particionados con relación a los dos bloques de variables.

Con respecto al segundo término, calculamos la esperanza del segundo término, observemos primero que

$$E\left[tr\left(V^{-1}\sum_{i=1}^{n} y_i\, y_i'\right)\right] = tr\left[E\left(V^{-1}\sum_{i=1}^{n} y_i\, y_i'\right)\right] = tr[V^{-1}\sum_{i=1}^{n} E(y_i\, y_i')]$$

Por tanto tenemos que obtener las esperanzas $E(y_i\, y_i'|Y, \hat{\theta}^{(i)})$ y consideramos:

Que en el caso del vector y_i se observa parcialmente y no se conoce los valores de y_{2i} pero si los de y_{1i}, entonces se obtendrá la relación

$$E\left(y_{2i}\, y_{2i}{}' \middle| Y, \hat{\theta}^{(i)}\right) = E\left(y_{2i}\, y_{2i}{}' \middle| y_{1i}, \hat{\theta}^{(i)}\right) = \hat{V}_{22|1}^{(i)} + \hat{y}_{2i.1}^{(i)}\, \hat{y}_{2i.1}^{(i)\,\prime}$$

Donde $\hat{V}_{22|1}^{(i)}$ es la matriz de varianzas de la variable y_{2i} dado y_{1i} y es equivalente a

$$\hat{V}_{22|1}^{(i)} = \hat{V}_{22}^{(i)} - \frac{\hat{V}_{21}^{(i)}}{\hat{V}_{11}^{(i)}}\, \hat{V}_{12}^{(i)}$$

Paso M: es la etapa final, se reemplazará las estimaciones obtenidas en las expresiones $E\left(y_i \middle| Y, \hat{\theta}^{(i)}\right)$ y $E\left(y_{2i}\, y_{2i}{}' \middle| Y, \hat{\theta}^{(i)}\right)$ en la función de verosimilitud y se calcula los nuevos estimadores de máxima verosimilitud, que vendrán dados por:

$$\hat{\mu}^{(i+1)} = \frac{\sum_{i=1}^{n} \hat{y}_i^{(i)}}{n}$$

donde en los casos en que $\hat{y}_i^{(i)}$ tenga valores observados no se modifican y aquellos que presenten ausencia de valor se sustituyen por las esperanzas condicionales obtenidas en $E\left(y_i \middle| Y, \hat{\theta}^{(i)}\right)$. La estimación de la matriz de covarianza esta expresada por:

$$\hat{V}^{(i+1)} = \frac{\sum_{i=1}^{n} E\left(y_{2i}\, y_{2i}{}' \middle| Y, \hat{\theta}^{(i)}\right)}{n} - \hat{\mu}^{(i+1)} \hat{\mu}^{(i+1)\prime}$$

Donde en los casos donde haya valores faltantes se sustituirán por las expresiones obtenidas de $E\left(y_{2i}\, y_{2i}{}' \middle| Y, \hat{\theta}^{(i)}\right)$.

Una vez que se tenga los valores estimados en el paso anterior denominado paso M se volverá a repetir el paso E haciendo $\hat{\mu}^{(i)} = \hat{\mu}^{(i+1)}$ y $\hat{V}^{(i)} = \hat{V}^{(i+1)}$ y dicho proceso finalizará cuando la diferencia entre los valores de los parámetros estimados sea un valor pequeño.

CAPÍTULO IV IMPUTACIÓN MÚLTIPLE DE DATOS FALTANTES

4.1 LA IMPUTACIÓN MÚLTIPLE

El método de Imputación Múltiple surgió como una alternativa moderna frente a los métodos simples que se utilizaron y entre los que se encontraron deficiencias que ocasionaban la obtención de parámetros estimados lejanos de la realidad. Es así, y gracias al gran aporte de Dempster y Laird en el año 1977 con la formalización del algoritmo EM y con la incursión del uso de métodos computacionales, se desarrolló el método de Imputación Múltiple.

Rubin (1987) introdujo el concepto de imputación múltiple planteando un método para promediar las estimaciones obtenidas de m conjuntos de datos imputados, imputaciones que son obtenidas a través de un proceso similar a la regresión estocástica pero un número repetido de veces, generando así los m conjuntos imputados de los cuales se obtendrán las estimaciones que finalmente serán promediadas para obtener una estimación global.

4.2 PROCEDIMIENTOS DE LA IMPUTACIÓN MÚLTIPLE

Los fundamentos bayesianos en los que se basa la Imputación Múltiple acondicionan el proceso de imputación que es el proceso más importante del conjunto de pasos en el que consiste el presente método. De manera general, la Imputación Múltiple consiste en tres procesos, el proceso de imputación, proceso de análisis y finalmente el proceso de combinación.

4.2.1 FASE DE IMPUTACIÓN

El siguiente proceso consiste en sustituir o imputar cada valor perdido por un conjunto de $m > 1$ valores independientes, desde la perspectiva bayesiana los valores imputados son valores aleatorios que serán obtenidos a través de la distribución condicional o distribución predictiva posterior de los datos faltantes $p\left(Y_{fal}\middle|Y_{obs}, \theta^{*}\right)$ donde θ^{*} es obtenida de la distribución posterior de los valores observados $p(\theta|Y_{obs})$.

La imputación múltiple genera m conjuntos con diferentes valores imputados, para este procedimiento se usa un algoritmo iterativo parecido al del algoritmo EM. La imputación adiciona valores aleatorios que ocasionan que se obtengan valores imputados con incertidumbre a través de los valores observados Y_{obs}. Los valores imputados obtenidos son valores aleatorios obtenidos a través de la distribución predictiva posterior de los datos faltantes $p(Y_{fal}|Y_{obs}, \theta)$ que depende de los valores observados y de los estimadores obtenidos respectivamente.

$$Y_t^{*} \sim P(Y_{fal} \ |Y_{obs}, \theta_{t-1}^{*})$$

Donde:

Y_t^{*} : denota los valores imputados en el paso (t) de Imputación.

Y_{fal} : denota los valores no observados o faltantes

θ_{t-1}^{*}: denota los estimadores obtenidos del paso (t-1) Posterior.

4.2.2 FASE DE ANÁLISIS

El siguiente proceso consiste en el análisis de los m conjuntos de datos completos que fueron obtenidos en la etapa previa, estos m conjuntos serán analizados y tratados como

si cada uno se tratara de los conjuntos de datos reales completos a los cuales se les aplicará los m análisis estadísticos de acuerdo al objetivo del investigador, es decir si entre los objetivos del investigador esta estimar una ecuación múltiple de regresión, entonces se ajustarán con los m conjuntos de datos m ecuaciones múltiple de regresión.

En este proceso finalmente se obtendrán m diferentes estimaciones de cada parámetro los cuales serán insesgados si presenta un mecanismo de datos faltantes MAR.

4.2.3 FASE DE COMBINACIÓN

En el tercer y último proceso, se procede a combinar las m estimaciones de los parámetros obtenidos y convertirlos en un solo punto estimado. Rubin (1987) definió el punto estimado de imputación múltiple como la media aritmética de las m estimaciones

$$\bar{\theta} = \frac{1}{m}\sum_{t=1}^{m} \hat{\theta}_t$$

donde $\hat{\theta}_t$ es el parámetro estimado de los conjuntos de datos completos y $\bar{\theta}$ es la estimación global. Las varianzas estimadas obtenidas también son combinadas y consisten dos componentes que toman en cuenta la variabilidad dentro y a través de cada conjunto de datos.

La matriz de covarianza dentro de la imputación es la media aritmética de las m matrices de covarianzas obtenidas

$$V_W = \frac{1}{m}\sum_{t=1}^{m} var(\hat{\theta}_t)$$

Donde V_w denota el promedio de las matrices de covarianzas dentro de la imputación y $var(\hat{\theta}_t)$ es el parámetro matriz de covarianza del conjunto de datos t, por lo tanto V_W será el valor que estimará la matriz de covarianza que hubiera resultado de haber obtenido los datos completos.

La matriz de covarianza entre imputaciones obtiene la variabilidad de los parámetros estimados a través de los m conjuntos de datos

$$V_B = \frac{1}{m-1}\sum_{t=1}^{m}(\hat{\theta}_t - \bar{\theta})(\hat{\theta}_t - \bar{\theta})'$$

donde V_B denota la matriz de covarianza entre imputaciones, $\hat{\theta}_t$ el parámetro estimado del conjunto de data t y $\bar{\theta}$ es el vector de puntos estimado promedio.

Por tanto la matriz de covarianza parámetro total obtenida T, es la suma de ambos componentes con un factor de corrección adicional

$$T = V_W + (1 + \frac{1}{m})V_B$$

La raíz cuadrada $\sqrt{T}$ es el error estándar total asociado a $\bar{\theta}$.

Si el tamaño de la data completa es grande y el número de imputaciones m es pequeña las pruebas de hipótesis y los intervalos de confianza están basados en una distribución de referencia t- Student

$$\frac{(\bar{\theta} - \theta)}{T} \sim t_v$$

donde los grados de libertad están dados por

$$V = (m-1)[1 + \frac{V_W}{(1+m^{-1})V_B}]^2$$

Por tanto un intervalo estimado al $100\,(1-\alpha)\%$ de Q es $\bar{Q} \pm t_{v,1-\alpha/2}\sqrt{T}$.

Por tanto los grados de libertad V no sólo dependen de m también de la expresión radio

$$r = \frac{(1+m^{-1})V_B}{V_W}$$

Expresión a la que Rubin (1987) denominó como el incremento relativo de la varianza debido a los valores faltantes. Cuando m sea lo suficientemente grande o r sea pequeño los grados de libertad serán grandes y $\frac{(\bar{\theta}-\theta)}{T} \sim t_v$ será aproximadamente normal.

De la cual se obtiene una fracción estimada de data faltante,

$$\gamma = \frac{(r+2)/(V+3)}{r+1}$$

Ambas expresiones son diagnóstico sutiles para caracterizar que tan fuerte la estimación de θ puede estar influenciada por los datos faltantes.

4.3 CONSIDERACIONES DE LA IMPUTACIÓN MÚLTIPLE

Es importante tomar en cuenta ciertas consideraciones con respecto al uso de la Imputación Múltiple de datos faltantes, y en Schafer (1999), se esclarecieron algunas dudas que se sostuvieron con respecto al presente método:

- Aún no era completamente claro cuál de los métodos usar, Listwise Deletion o la Imputación Múltiple, ya que la mayoría de usuarios optaban por utilizar el primer método debido a su fácil aplicación y entendimiento. Esta duda que se plantea se espera sea despejada con las comparaciones de los resultados obtenidos.

- La práctica ha demostrado que en situaciones en que la tasa de no respuesta es baja se sugiere aplicar el método de Imputación Múltiple, especialmente cuando en el análisis se utilizan técnicas multivariadas.

- En cuanto al número de imputaciones a realizarse el presente método es capaz de generar resultados robustos hasta con un número pequeño de iteraciones. De acuerdo al autor Rubin (1987, p.114), la eficiencia relativa de m imputaciones es medida como $(1 + \lambda/m) - 1$, en donde λ es la tasa de registros sin información. También señala que, para tasas de respuesta inusualmente altas sólo se requiere generar entre 5 y 10 imputaciones.

- Con respecto al parecido del procedimiento de estimación por máxima verosimilitud, ambas metodologías aplican métodos numéricos y simulaciones de Monte Carlo, y se demuestra que para tamaños de muestra grandes generan resultados similares.

- Schafer (1999), afirma que en la aplicación del paradigma de la Imputación Múltiple no es necesario suponer que el patrón de observaciones faltantes debe ser ignorado, y señala que los procedimientos desarrollados por Rubin pueden ser aplicados a cualquier tipo de modelos de simulación. También reconoce que el tema está en desarrollo, y que resolver satisfactoriamente este tipo de inquietudes aún representa un importante desafío.

CAPÍTULO V APLICACIONES

5.1 INTRODUCCIÓN

Uno de los modelos de probabilidad más comunes para datos multivariados continuos es el de la distribución normal, la cual forma parte como supuesto de muchos métodos estándares como análisis factorial, componentes principales, análisis discriminante, etc. Debido a que diferentes métodos estadísticos utilizan supuestos de normalidad es común que se opte por buscar técnicas de inferencia para modelos de data incompleta con esta distribución.

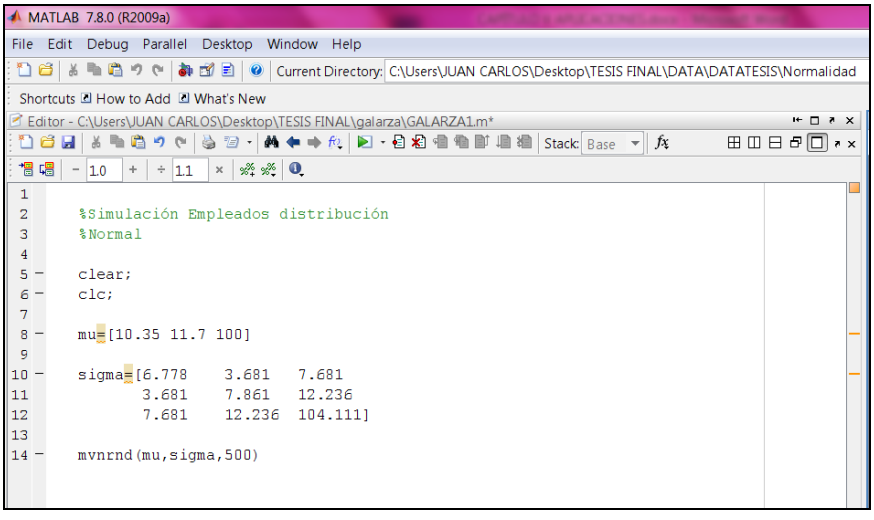
Desde un inicio se pensó en la búsqueda de un conjunto de base de datos con una distribución normal multivariada de muestra grande, para poder aplicar ambos métodos, debido a la dificultad de obtener un conjunto de datos que cumpla con el supuesto de normalidad es que se pensó utilizar herramientas computacionales como la simulación, con el objetivo de generar un muestra que cumpla con las características necesarias. Una vez obtenido el conjunto de datos se obtendrán los parámetros estimados, es decir los parámetros de la muestra simulada sin ningún caso faltante. Con el objetivo de poner en práctica las técnicas de inferencias precisadas en los capítulos anteriores (el algoritmo EM e Imputación múltiple), se ocasionará la pérdida de algunos valores generando datos faltantes para aplicar posteriormente el algoritmo EM y la imputación Múltiple. Es importante precisar que se aplicarán ambas metodologías con el afán de comparar los resultados y a la vez corroborar si distan con los valores obtenidos con la data completa.

5.2 SIMULACIÓN DE LA MUESTRA

Para ilustrar la efectividad de ambos métodos se utilizaron herramientas computacionales como la simulación haciendo uso de algoritmos y pseudocódigos en MATLAB. La data generada se tomó como referencia del libro de Enders, C. K. (2010) que recopiló la información del artículo de investigación de Thomas A. Wright y Douglas G. Bonett (2007) en el cual examinaron diversas características laborales que influyen en la rotación o cambio de empleados. De la data se sacaron las siguientes variables:

Y_1: Bienestar Psicológico, ***Y_2: Rendimiento Laboral,*** ***Y_3: IQ***

Se utilizó la simulación para generar los puntajes de $N = 500$ empleados con las tres características en mención con un modelo de distribución normal, utilizando la función $mvnrnd$ que es la que generará el conjunto de vectores aleatorios con distribución normal multivariada con media mu y covarianza $sigma$, valores que son especificados.



Con respecto a la matriz de covarianza es importante recalcar una particularidad con respecto al algoritmo EM, ya que esta metodología trabaja con inversas de la matriz de covarianzas V^{-1} y por tanto debe tenerse cuidado en cuanto al plantear un matriz que presente la característica de singularidad, pues estas matrices cuadradas tiene determinante cero, determinante que se utiliza para implementar el paso E del algoritmo EM. Por tanto si esta matriz de covarianzas no es definida positiva el resultado sería imaginario y en consecuencia el algoritmo fallaría. Para la simulación es importante declarar una matriz de covarianzas que cumpla con ser semi-definida positiva, y así evitar el problema en el cálculo de los estimadores en el paso-M, en el cuál es necesario el cálculo de la matriz inversa.

5.3 NORMALIDAD DE LA MUESTRA

Será importante verificar el supuesto de normalidad en el sentido univariado y multivariado, ya que con sólo la condición de distribución normal univariado es necesaria mas no suficiente para comprobar que sea una distribución normal multivariante. Debido a una rápida evaluación de manera directa se comprobará si en conjunto las variables tienen un modelo de distribución normal, haciendo uso de los ***Test de Simetría y Kurtosis***, planteados por Mardia (1970).

Para el primer conjunto de datos creados se obtiene se plantea las siguientes hipótesis:

$Ho: La\ distribución\ de\ Y_{nxp}\ es\ simétrica.$

Obtenemos el coeficiente de simetría $n\frac{A_p}{6}$ donde $A_p = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^3$ y

$f = \frac{1}{6}\, p\,(p+1)(p+2)$, si $n\frac{A_p}{6} > \chi_f^2$ entonces se rechaza la hipótesis nula.

El coeficiente de asimetría es: $n\frac{A_p}{6}$ = 6,0582, f = 10, χ_f^2 = 18,3070

Con los valores obtenidos por tanto no se rechaza la hipótesis nula, quedando comprobado que la distribución es simétrica.

$Ho: La\ distribución\ de\ Y_{nxp}\ es\ mesocúrtica$

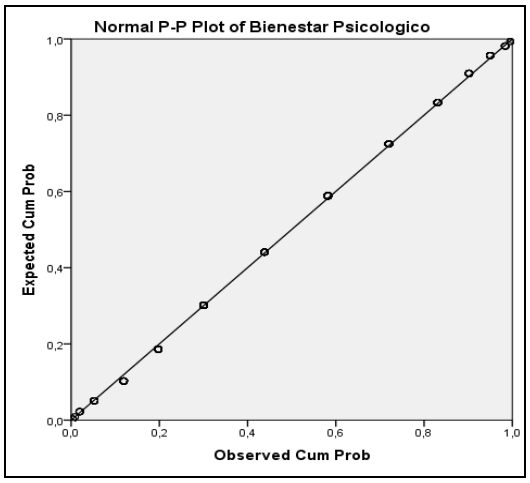
Obtenemos el coeficiente de kurtosis, $K_p = \frac{1}{n}\sum_{i=1}^{n} d_{ii}^2$, $K_p \sim N\left[p\,(p+2); \frac{8\,p\,(p+2)}{n}\right]$

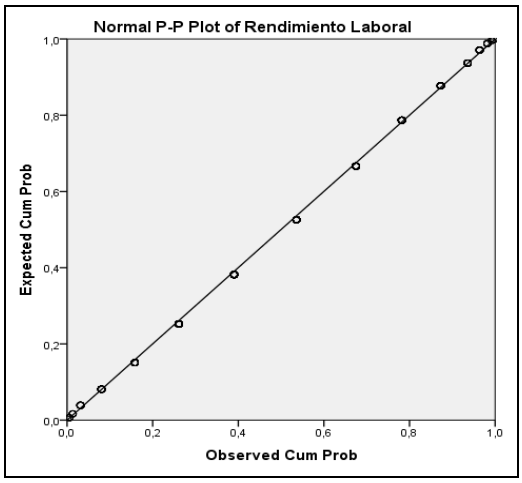
$K_p^* = \frac{K_p - p(p+2)}{\sqrt{\frac{8p(p+2)}{n}}} \sim N(0,1)$, si $K_p^* > Z_{\frac{\alpha}{2}}$ o $K_p^* < -Z_{\frac{\alpha}{2}}$ no se rechaza la hipótesis nula.

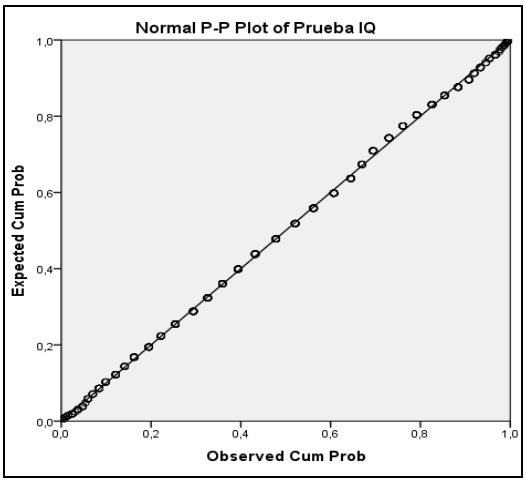
El coeficiente de kurtosis es, K_p^*= -0,4606, $Z_{\frac{\alpha}{2}}$ = 1,6449

Con los valores obtenidos por tanto no se rechaza la hipótesis nula, quedando comprobado que la distribución es mesocúrtica.

Con ambas pruebas de hipótesis concluyendo que la matriz de datos Y_{nxp} es simétrica y mesocúrtica, podemos concluir y comprobar que el conjunto de datos tiene una distribución normal multivariada. Para concluir realizamos pruebas gráficas P-P univariadas de cada una de las variables en cuestión, quedando comprobado que cada una de las variables muestran una distribución normal univariada.







5.4 MUESTRA ORIGINAL DE DATOS

En el análisis exploratorio se obtienen los estadísticos de la muestra simulada de los $N = 500$ empleados, en los siguientes cuadros obtenidos con el programa SPSS v.20 se muestran los estadísticos descriptivos, la matriz de varianza, covarianza y correlación de la data original en la que no hay presencia de información faltante, los que serán objeto en una posterior comparación.

Estadística Descriptiva				
	Bienestar Psicológico	Rendimiento Laboral	Prueba IQ	N válido(listwise)
N	500	500	500	500
Media	10.40	11.82	100.54	

Matriz de Covarianza			
	Bienestar Psicológico	Rendimiento Laboral	Prueba IQ
Bienestar Psicológico	7.2184	3.6597	8.5691
Rendimiento Laboral	3.6597	7.4800	11.3858
Prueba IQ	8.5691	11.3858	98.1768

Matriz de Correlación			
	Bienestar Psicológico	Rendimiento Laboral	Prueba IQ
Bienestar Psicológico	1.0000	0.4981	0.3219
Rendimiento Laboral	0.4981	1.0000	0.4202
Prueba IQ	0.3219	0.4202	1.0000

Se ocasiona la pérdida de ciertos valores de la data completa con el objetivo de crear un conjunto de datos incompletos para poder así hacer uso de las metodologías precisadas, tomando en cuenta el mecanismo de pérdida de los datos.

5.5 APLICACIÓN DE LA METODOLOGÍA

Para la presente aplicación se tomaron en cuenta los dos mecanismos de pérdida de datos más comunes **MAR** y **MCAR**. Es importante precisar que el mecanismo de pérdida de datos modelado para las metodologías EM e Imputación Múltiple es **MAR** - perdidos al azar, que refiere que la probabilidad de la ausencia de datos en una variable está relacionada con alguna otra u otras variables observables. Si bien es cierto, ambas metodologías trabajan con este tipo de mecanismo de datos, en la actualidad no existe prueba estadística para comprobar cuando una muestra con valores faltantes presentan un mecanismo **MAR** generando un problema práctico en la aplicación de ambas metodologías.

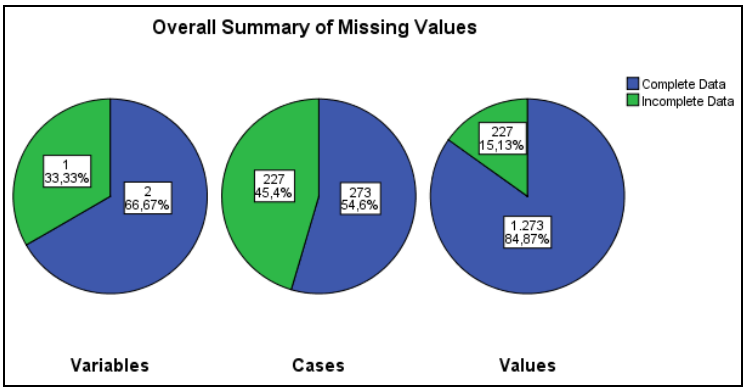
Ante la ausencia de una prueba que certifique este mecanismo, la práctica ha mostrado que en muestras grandes de datos que presentan este inconveniente, la aplicación del algoritmo EM y la IM actúan obteniendo aún mejores estimadores que las técnicas de imputación simple precisados en capítulos precedentes ya que estos métodos maximizan el poder estadístico haciendo uso de información observada. Por eso, con el objetivo de probar la eficacia de los métodos aun frente al mecanismo **MCAR** (mecanismo que asume que la pérdida de los datos no está relacionada con otras variables observables ni consigo misma); es que se presenta un caso práctico para probar la eficacia de las metodologías, un punto a favor con este mecanismo es la existencia de una prueba estadística Litres MCAR que presenta una comparación de medias en subgrupos. Por tanto se trabajaron con tres casos en la cual se alterará la data ocasionando datos faltantes.

5.6 CASO 1: DATOS FALTANTES EN LA PRUEBA DE RENDIMIENTO LABORAL

Con datos perdidos presente en el rendimiento laboral y con un mecanismo de pérdida MAR**.** La pérdida de información en el Rendimiento Laboral de algunos empleados ha sido ocasionado por el bajo puntaje que obtuvieron en la prueba de IQ, aquellos empleados que no sobrepasaron los 100 puntos en la prueba IQ, fueron eliminados del proceso de selección y por consecuente no se registró el Rendimiento Laboral, por lo tanto planteamos que la ausencia de datos en el Rendimiento Laboral está relacionado con los resultados en la prueba IQ.

5.6.1 ESTADÍSTICOS DESCRIPTIVOS DE LOS DATOS FALTANTES

En el SPSS v20 se puede obtener un análisis descriptivo a cerca de los valores faltantes, el cual es necesario realizar para ver el comportamiento de los valores perdidos frente a la muestra observada. De esta manera obtenemos los siguientes gráficos y tablas:

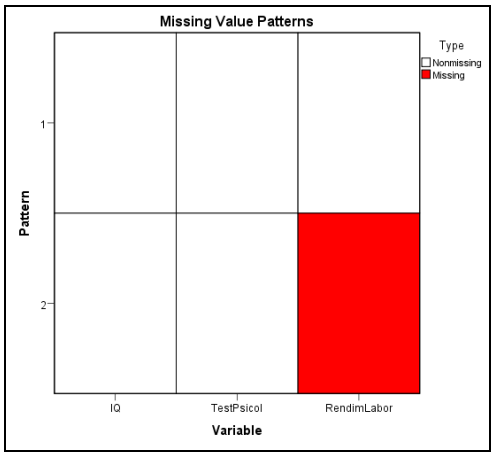


De la presente tabla podemos observar de manera gráfica los porcentajes de valores perdidos según:

- Variables: sólo se presenta una sola variable que cuenta con información faltante, representa el 33,3 por ciento de las variables.
- Casos: se presenta 227 casos o unidades informativos con valores perdidos, es decir 45,4 por ciento de los informantes al menos tiene un valor faltante entre sus respuestas.
- Valores: del número total de casos, es decir de 1500 valores en la data solo un 15,13 por ciento es representado como valor faltante.

5.6.2 MATRIZ DE PATRÓN DE DATOS FALTANTES

El patrón de datos faltante que se presenta con la matriz respectiva es de la forma univariada, ya que la ausencia de datos se centra sólo en los puntajes del Rendimiento Laboral representado en 227 empleados (45,4 por ciento).



Resumen de Variables					
	Perdidos		N Válido	Media	Desviación Estándar
	N	Porcentaje			
Rendimiento Laboral	227	45.4%	273	12.62	2.571

5.6.3 MECANISMO DE DATOS FALTANTES

Para la identificación del mecanismo de datos faltantes de la muestra se utilizaron una serie de pruebas t independiente que realizan comparaciones entre los subgrupos de datos faltantes para examinar si difieren entre los grupos de medias obtenidas, esto puede ayudar a identificar si la ausencia de los puntajes del rendimiento laboral se deba a alguna otra variable presente en la muestra y así finalmente verificar si el mecanismo de datos faltantes es MCAR.

Se usó el programa SPSS v20 para obtener dichos valores entre los subgrupos, esta prueba forma subgrupos de cada variable observable, en este caso de los puntajes de la prueba IQ y el bienestar psicológico, tomando en cuenta si se tiene datos con respecto al rendimiento laboral.

Prueba t de varianzas separadas				
		Bienestar Psicológico	Rendimiento Laboral	IQ
Rendimiento Laboral	t	6.1		29.5
	df	480.7		491.0
	P(2-colas)	0.0		0.0
	# Presente	273.0	273.0	273.0
	# Perdidos	227.0	0.0	227.0
	Media(Presente)	11.0	12.6	107.7
	Media(Perdidos)	9.6		91.9

Los casos con información perdida y completa del rendimiento laboral obtuvieron un puntaje del bienestar psicológico promedio de 9,6 y 11,0 respectivamente, y una prueba T-Student - Welch $t(480{,}7) = 6{,}1\ p < 0{,}05$. Con respecto a los casos con información perdida y completa del rendimiento laboral se obtuvo puntajes IQ promedios 91,9 y 107,7 y de la misma manera aplicando la prueba T- Student – Welch $t(491) = 29{,}5\ p < 0{,}05$.

En ambos casos resultaron que las diferencias de los promedios de los subgrupos con ausencia y presencia de información distan significativamente lo que nos indicaría que los valores faltantes en el rendimiento laboral no tienen un mecanismo de datos faltantes MCAR, por tanto pueden ser que hayan sido perdidos al azar – MAR o no perdidos al azar MNAR, esta conclusión coincide con la alteración que se realizó a los puntajes del rendimiento laboral ocasionando un mecanismo MAR (los empleados que no

sobrepasaron los 100 puntos en la prueba IQ no se les tomó la prueba de rendimiento laboral) .

Otro de las maneras de comprobar que el mecanismo de pérdida de la presente muestra no es MCAR, es utilizando la prueba Little's MCAR, presente en el SPSS v20.0 que a continuación presentamos:

La prueba de hipótesis Little's MCAR, nos ayudará a comprobar si el conjunto de datos tiene un patrón de pérdida completamente al azar es decir MCAR.

H_0: Los datos están completamente perdidos al azar (MCAR).

H_1 : Los datos no están completamente perdidos al azar.

$$Chi - cuadrado = 315{,}935 \quad DF = 2 \quad p - valor = 0{,}000$$

Dado que el p-valor es 0,00 y dicho valor es < 0,05, podemos concluir que los datos no están completamente perdidos al azar, esto corrobora la afirmación entablada en la prueba T-Student – Welch anterior.

El test de Little's MCAR nos confirma únicamente que el mecanismo de datos faltantes no es MCAR, mas no nos confirma que sea MAR, por eso es importante analizar otros aspectos descriptivos para de estar manera obtener indicios si la ausencia de información se deba a la presencia de alguna otra variable presente en la muestra, como es en el presente caso.

5.6.4 APLICACIÓN DEL ALGORITMO EM E IMPUTACIÓN MÚTIPLE DE DATOS

La primera metodología a utilizar es el algoritmo EM para estimar medias, varianzas, matriz de covarianzas y de correlación de los datos. Para tales fines se diseñó medidas de desempeño del método EM y la imputación múltiple mediante código fuente en MATLAB para ejecutar los procesos, de esta manera posteriormente poder comparar los estimadores obtenidos y analizar que método obtuvo las mejores estimaciones con respecto a la muestra original.

El algoritmo EM desarrollado obtiene de los datos observables las estimaciones de medias, varianzas y covarianzas, valores que posteriormente utiliza para estimar los valores faltantes. Una vez que se obtenga una data completa con los valores sustituidos, es que se maximiza obteniéndose nuevamente las estimaciones de medias, varianzas y covarianzas, este procedimiento se realiza de manera sucesiva hasta que la diferencia entre parámetros no sea significativa.

Iniciamos el proceso de algoritmo EM, haciendo correr el programa en el cual se especifica el número de iteraciones a correr y el criterio de convergencia. El programa nos arroja como parte inicial un resumen de los datos ingresados, especificando el vector de medias y la matriz de covarianza de la muestra incompleta, valores que toma como parámetros iniciales. Se generó tres pruebas con 20, 30, 50 iteraciones respectivamente, el programa nos arroja resultados óptimos de convergencia con una cantidad de 50 iteraciones.

Posteriormente se utilizaron los parámetros estimados en el algoritmo EM tomándolo como punto inicial para iniciar el método de imputación múltiple de datos. Este método estimará los datos que faltan, para luego en base al enfoque Bayesiano crear una distribución de probabilidad de los valores de los parámetros. En el programa ejecutado en MATLAB, se especifica el número de imputaciones a ejecutar, se realizó pruebas de 5 a 10 imputaciones siendo esta última la cantidad presentada para el análisis correspondiente, de estas 10 imputaciones se procedió a obtener los estimadores globales correspondientes al proceso de combinación de los promedios y varianzas estimadas en cada uno de los conjuntos imputados, presentándose así finalmente las estimaciones obtenidas con referencia a estos métodos, descritos en los cuadros correspondientes.

5.6.5 COMPARACIÓN DE RESULTADOS

A continuación mostramos los parámetros estimados en los distintos casos, cuando se tuvo la muestra original completa, cuando se optó por usar el método Listwise Deletion, y con la muestra incompleta (por mecanismo MAR) aplicando el algoritmo EM y el método de Imputación Múltiple para datos faltantes.

Estimación	Parámetro Poblacional	Técnica de Data Faltante		
		Listwise Deletion	Algoritmo EM	Imputación Múltiple
		Simulación MAR		
BP Media	10.400	10.500	10.400	10.400
RL Media	11.824	12.620	11.702	11.699
IQ Media	100.540	100.540	100.540	100.540
BP Varianza	7.218	6.700	7.204	7.282
RL Varianza	7.480	6.612	7.589	9.066
IQ Varianza	98.177	38.238	97.980	98.985
BP - RL Covarianza	3.660	2.978	3.697	3.942
BP - IQ Covarianza	8.569	2.899	8.552	8.784
RL - IQ Covarianza	11.386	4.662	12.395	12.774
BP - RL Correlación	0.498	0.447	0.500	
BP - IQ Correlación	0.322	0.181	0.322	
RL - IQ Correlación	0.420	0.293	0.454	

Como podemos visualizar en la tabla respectiva, realizamos una comparación de los estimadores obtenidos de la muestra original, de la muestra con los valores faltantes con mecanismo de perdida MAR utilizando el Listwise Deletion, y los estimadores obtenidos usando el algoritmo EM y la Imputación Múltiple de datos. Con respecto a la media el uso del Listwise Deletion tiende a distorsionar las estimaciones originales, sobrestimando dicho valor mientras que las estimaciones obtenidas con el algoritmo EM y la IM se ven más cercanos a los valores reales.

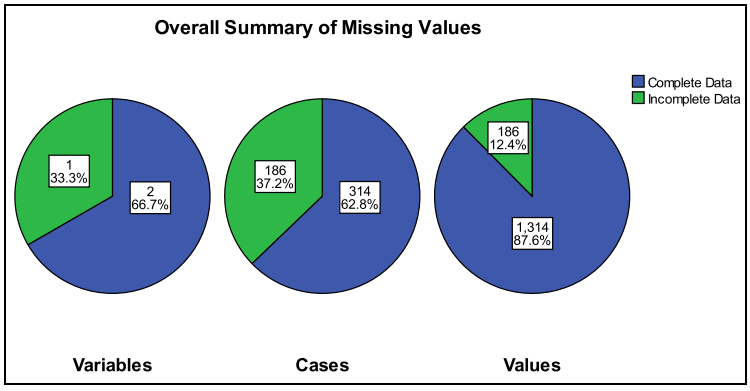
En cuanto a las covarianzas en algunos casos el algoritmo EM y la IM obtienen estimadores con cierta sobrestimación pero que no representan una diferencia notable con respecto a las covarianza real obtenida de la muestra original, mientras que la varianza de la prueba IQ se ve afectada en el caso del uso de Listwise Deletion debido a la cantidad de datos perdidos, en ese caso el algoritmo EM y la IM si han logrado obtener estimadores más cercanos al valor real.

5.7 CASO 2: DATOS FALTANTES EN LA PRUEBA DE BIENESTAR PSICOLÓGICO

Con datos perdidos presente en la prueba de bienestar psicológico con un mecanismo de perdida MCAR. La pérdida de información en esta prueba ha sido ocasionado de manera aleatoria utilizando la herramienta del Excel generación de números aleatorios que nos originó la posición de los valores faltantes, por consecuente en este caso no estamos precisando que la ausencia de datos se deba a algún tipo de relación entre variables presentes en la muestra por eso especificamos que este tipo de mecanismo sea considerado como MCAR.

5.7.1 ESTADÍSTICOS DESCRIPTIVOS DE LOS DATOS FALTANTES

Observamos el comportamiento de los valores perdidos de nuestro segundo caso. De esta manera obtenemos los siguientes gráficos y tablas:



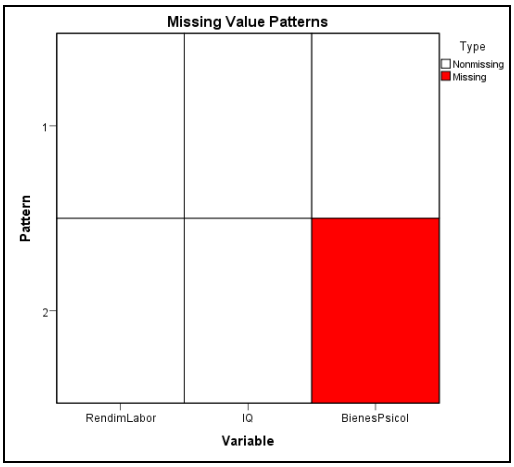
De la presente tabla podemos observar de manera gráfica los porcentajes de valores perdidos según:

- Variables: sólo se presenta una sola variable que cuenta con información faltante, representa el 33,3 por ciento de las variables.

- Casos: se presenta 186 casos o unidades informativos con valores perdidos, es decir 37,2 por ciento de los informantes al menos tiene un valor faltante entre sus respuestas.
- Valores: del número total de casos, es decir de 1500 valores en la data solo un 12,4 por ciento es representado como valor faltante.

5.7.2 MATRIZ DE PATRÓN DE DATOS FALTANTES

El patrón de datos faltante que se presenta con la matriz respectiva es de la forma univariada, ya que la ausencia de datos se centra solo en los puntajes del bienestar psicológico representado en 186 empleados (37,2 por ciento).



Resumen de Variables					
	Perdidos		N válido	Media	Desviación Estándar
	N	Porcentaje			
Bienestar Psicológico	186	37.2%	314	10.516	2.639

5.7.3 MECANISMO DE DATOS FALTANTES

Prueba t de varianzas separadas				
		Bienestar Psicológico	Rendimiento Laboral	IQ
Bienestar Psicológico	t		-.1	-.2
	df		362.3	380.7
	P(2-colas)		.928	.878
	# Presente	314	314	314
	# Perdidos	0	186	186
	Media(Presente	10.52	11.82	100.49
	Media(Perdidos		11.84	100.63

De la misma manera se usó el programa SPSS v20 para obtener los promedios entre los subgrupos, en este caso de los puntajes de la prueba IQ y el rendimiento laboral, tomando en cuenta si hay presencia o no de información con respecto al bienestar psicológico.

Los casos con información perdida y completa del bienestar psicológico obtuvieron un puntaje del rendimiento laboral promedio de 11,84 y 11,82 respectivamente, y una prueba T-Student - Welch $t(362{,}3) = -0{,}1\ p > 0{,}05$. Con respecto a los casos con información perdida y completa del bienestar psicológico se obtuvieron puntajes IQ promedios 100,63 y 100,49 y de la misma manera aplicando la prueba T- Student – Welch $t(380{,}7) = -0{,}2\ p > 0{,}05$.

En ambos casos resultaron que las diferencias de los promedios de los subgrupos con ausencia y presencia de información no distan significativamente lo que nos haría pensar que la ausencia de los puntajes en el bienestar psicológico se haya debido por algún motivo referido a los puntajes IQ y al rendimiento laboral de los empleados, es decir que la muestra con información faltante tiene un mecanismo de datos faltantes MCAR.

A continuación aplicamos la prueba de hipótesis Little's MCAR, que nos ayudará a comprobar si el conjunto de datos tiene un patrón de perdida completamente al azar (MCAR).

H_0: Los datos están completamente perdidos al azar (MCAR).

H_1 : Los datos no están completamente perdidos al azar.

$$Chi - Cuadrado = 0{,}025 \qquad DF = 2 \qquad p - valor = 0{,}988$$

Dado que el p-valor no es significativo ya que es superior a 0,05, no rechazamos la hipótesis nula y podemos concluir que los datos están completamente perdidos al azar (MCAR), esto corrobora la afirmación entablada en la prueba anterior.

5.7.4 APLICACIÓN DEL ALGORITMO EM E IMPUTACIÓN MÚTIPLE DE DATOS

Nuevamente hacemos uso del código fuente en MATLAB para ejecutar las metodologías y luego comparar los estimadores obtenidos, con el algoritmo EM solo se requirió de 7 iteraciones para que el programa nos arroje los estimadores respectivos, los cuales fueron utilizados como puntos iniciales para generar los 10 conjuntos de datos con el método de imputación múltiple.

5.7.5 COMPARACIÓN DE RESULTADOS

Los parámetros estimados en los distintos aspectos se comparan en el siguiente cuadro.

Estimación	Parámetro Poblacional	Técnica de Data Faltante		
		Listwise Deletion	Algoritmo EM	Imputación Múltiple
		Simulación MCAR		
BP Media	10.400	10.500	10.522	10.504
RL Media	11.824	11.824	11.824	11.824
IQ Media	100.540	100.540	100.540	100.540
BP Varianza	7.218	6.966	7.076	7.977
RL Varianza	7.480	7.007	7.465	7.487
IQ Varianza	98.177	96.525	97.980	98.761
BP-RL Covarianza	3.660	3.508	3.749	3.860
BP-IQ Covarianza	8.569	7.869	8.387	8.746
RL-IQ Covarianza	11.386	10.324	11.363	11.402
BP-RL Correlación	0.498	0.502	0.562	
BP-IQ Correlación	0.322	0.303	0.318	
RL-IQ Correlación	0.420	0.397	0.420	

En la tabla respectiva, ahora presentamos una muestra con los valores faltantes en los puntajes de la prueba de bienestar psicológico con mecanismo de perdida MCAR.

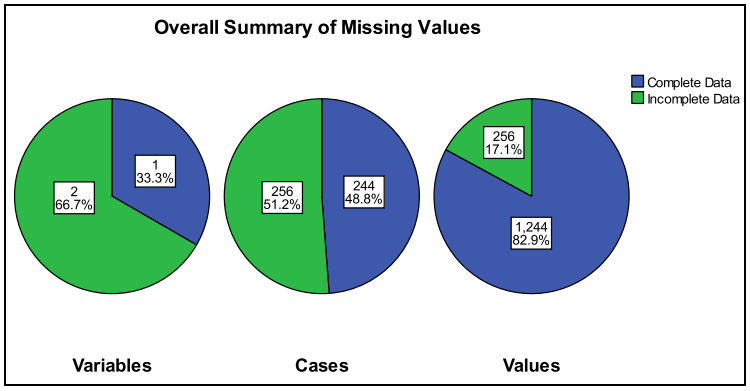
Las diferencias notables nuevamente se muestran en la matriz de covarianza con respecto a la variabilidad de los puntajes de la prueba IQ, ya que como podemos visualizar dicho valor original es 98,17, valor que es más cercano a la varianza estimada utilizando el algoritmo EM y la IM. Muy a pesar que el método Listwise Deletion obtuvo una varianza estimada de 96,52, los dos métodos planteados en esta tesis obtuvieron mejores valores; más cercanos al valor original. Este notable resultado se presenta aun en una muestra que tiene un mecanismo de pérdida de datos MCAR, no correspondiente a las metodologías del algoritmo EM y la IM.

5.8 CASO 3: DATOS FALTANTES EN LA PRUEBA DE RENDIMIENTO LABORAL Y BIENESTAR PSICOLÓGICO

Con datos perdidos en el Rendimiento Laboral y la prueba de bienestar psicológico con mecanismos MAR y MCAR de pérdida de datos respectivamente con este fin se espera comprobar la eficacia de la metodología aun en ambos mecanismos de pérdida de datos.

5.8.1 ESTADÍSTICOS DESCRIPTIVOS DE LOS DATOS FALTANTES

Observamos el comportamiento de los valores perdidos de nuestro segundo caso. De esta manera obtenemos los siguientes gráficos y tablas:



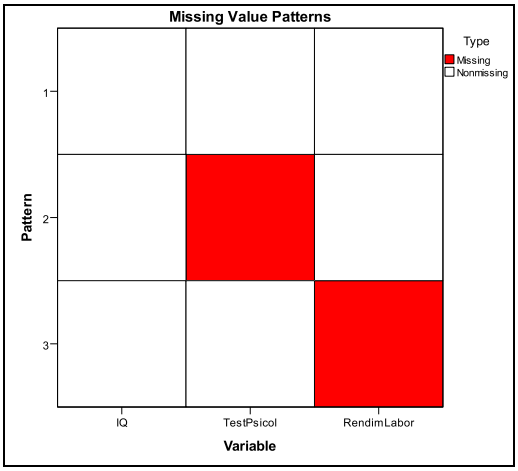
De la presente tabla podemos observar de manera gráfica los porcentajes de valores perdidos según:

- Variables: se presentan dos variables que cuenta con información faltante, representa el 66,7 por ciento de las variables.
- Casos: se presenta 256 casos o unidades informativas con valores perdidos, es decir 51,2 por ciento de los informantes al menos tiene un valor faltante entre sus respuestas.

- Valores: del número total de casos, es decir de 1500 valores en la data solo un 17,1 por ciento es representado como valor faltante.

5.8.2 MATRIZ DE PATRÓN DE DATOS FALTANTES

El patrón de datos faltante es monótono ya que la ausencia de los datos se centra en las variables Rendimiento Laboral y bienestar psicológico, la ausencia de datos está representada por un 45,4 y 5,8 por ciento respectivamente, los gráficos muestran la ausencia de información con los porcentajes respectivos.



Resumen de Variables					
	Perdidos		N válido	Media	Desviación Estándar
	N	Porcentaje			
Rendimiento Laboral	227	45.4%	273	12.62	2.571
Bienestar Psicológico	29	5.8%	471	10.35	2.729

5.8.3 MECANISMO DE DATOS FALTANTES

Prueba t de varianzas separadas				
		Bienestar Psicológico	Rendimiento Laboral	IQ
Bienestar Psicológico	t		.0	-4.6
	df		33.1	33.6
	P(2-colas)		.997	.000
	# Presente	471	244	471
	# Perdidos	0	29	29
	Media(Presente)	10.35	12.62	100.13
	Media(Perdidos)		12.62	107.24
Rendimiento Laboral	t	5.8		29.5
	df	468.1		491.0
	P(2-colas)	.000		.000
	# Presente	244	273	273
	# Perdidos	227	0	227
	Media(Presente)	11.03	12.62	107.72
	Media(Perdidos)	9.62		91.90

Los casos con información perdida y completa del bienestar psicológico obtuvieron un puntaje del rendimiento laboral promedio de 12,62 y 12,62 respectivamente, y una prueba T-Student - Welch $t(33{,}1) = 0{,}0\ p > 0{,}05$. Con respecto a los casos con información perdida y completa del bienestar psicológico se obtuvo puntajes IQ promedios 107,24 y 100,13 y de la misma manera aplicando la prueba T- Student – Welch $t(33{,}6) = -4{,}6\ p < 0{,}05$.

En ambos casos se obtuvieron resultaron diferentes, se tiene un mecanismo MCAR del bienestar psicológico con respecto a la ausencia y presencia de información en el rendimiento laboral, ya que las diferencias de los rendimientos promedios no distaban significativamente. Mientras que para los IQ promedios dichas diferencias eran estadísticamente significantes perfilándose un mecanismo de datos faltantes no MCAR.

De la misma manera los casos con información perdida y completa del rendimiento laboral obtuvieron un puntaje del bienestar psicológico promedio de 9,62 y 11,03 respectivamente, y una prueba T-Student - Welch $t(468{,}1) = 5{,}8 \; p < 0{,}05$. Mientras que se obtuvo puntajes IQ promedios 91,9 y 107,72 y de la misma manera aplicando la prueba T- Student – Welch $t(491) = 29{,}5 \; p < 0{,}05$.

En ambos casos se obtuvieron diferencias de promedios estadísticamente significativamente. Esto aduce que la ausencia de información en el rendimiento laboral no tiene un mecanismo MCAR.

5.8.4 APLICACIÓN DEL ALGORITMO EM E IMPUTACIÓN MÚTIPLE DE DATOS

Nuevamente hacemos uso del código fuente en MATLAB para ejecutar las metodologías y luego comparar los estimadores obtenidos. Con el algoritmo EM solo se nos requirió 50 iteraciones para que el programa nos arroje resultados óptimos, mientras que con esas estimaciones tomadas como puntos iniciales; se generaron 10 conjuntos de datos imputados haciendo uso del método de imputación múltiple.

5.8.5 COMPARACIÓN DE RESULTADOS

Los parámetros estimados en los distintos aspectos se comparan en el siguiente cuadro.

Estimación	Parámetro Poblacional	Técnica de Data Faltante		
		Listwise Deletion	Algoritmo EM	Imputación Múltiple
BP Media	10.400	10.350	10.390	10.390
RL Media	11.824	12.620	11.732	11.878
IQ Media	100.540	100.540	100.540	100.540
BP Varianza	7.218	7.159	7.464	7.540
RL Varianza	7.480	6.401	7.483	7.654
IQ Varianza	98.177	35.548	97.980	98.370
BP-RL Covarianza	3.660	2.766	3.470	3.417
BP-IQ Covarianza	8.569	2.764	8.665	8.729
RL-IQ Covarianza	11.386	3.796	12.076	10.757
BP-RL Correlación	0.498	0.409	0.464	
BP-IQ Correlación	0.322	0.173	0.320	
RL-IQ Correlación	0.420	0.252	0.445	

En este último caso donde ya presentamos una muestra con dos variables con ausencia de información, ambas con diferentes mecanismos de perdida MAR y MCAR respectivamente podemos notar valores diferentes.

La media obtenida para la muestra original fue 10,4 en el caso del bienestar psicológico y 11,824 para los puntajes del rendimiento laboral, si sólo hiciéramos uso de los valores observables obtendríamos valores 10,35 y 12,62 mientras que con el algoritmo EM y la IM 10,39, 11,732 y 10,39 11,878 respectivamente, estos valores tiene son más cercanos a la original con el método de Imputación Múltiple.

Con respecto a la matriz de covarianzas o varianzas dicha diferencia se hace notable en las varianzas obtenidas con el rendimiento laboral (7,48 en la muestra original, 6,401 usando el Listwise Deletion, 7,483 con el algoritmo EM y 7,654 con la IM) y la prueba IQ (98,117 en la muestra original, 35,548 usando el Listwise Deletion, 97,98 con el algoritmo EM y 98.37 con la IM), en la tabla el algoritmo EM obtuvo mejores resultados con respecto a la variabilidad del rendimiento laboral y la IM con respecto a la variabilidad de la prueba IQ. Mostrándonos que ambos métodos resultan ser eficaces en cuanto al uso de alguna metodología para tratar los valores faltantes dentro de un conjunto de datos.

CONCLUSIONES

1. Se pudo constatar que es necesario tener conciencia con respecto al manejo de un conjunto de datos que presenta ausencia de información, se requiere de cierto análisis previo, con el cual se indague la causa que originó dicho comportamiento.

2. Es importante que cuando se maneje conjuntos de datos grandes y haya presencia de información faltante, se debe tomar en cuenta utilizar algún tipo de metodología que ayude a lidiar con este inconveniente, ya que cuando existe una pérdida significativa los estadísticos obtenidos pueden reflejar resultados que distan mucho de la realidad.

3. Es necesario tomar en cuenta previamente, la cantidad de datos perdidos y el comportamiento de la variable que presenta la ausencia de información, ya que de esto dependerá la metodología de tratamiento de datos faltantes a utilizar. Pues si hablamos de una cantidad mínima, dicho problema pueda resolverse con métodos como el Listwise Deletion, en caso contrario este método no es el más recomendable.

4. Se describió las metodologías algoritmo EM y la Imputación Múltiple tanto así como su ejecución en el programa MATLAB donde se trabajó con un mismo conjunto de datos, planteando distintos casos de ausencia de información y obteniéndose finalmente las estimaciones necesarias para su posterior comparación.

5. Se obtuvo las estimaciones planteándose tres casos, con un patrón de pérdida de información univariada para los dos primeros casos, y con un patrón de pérdida general

para el tercero, también se planteó en el primer caso un mecanismo de perdida al azar (MAR), un mecanismo perdido completamente al azar (MCAR) para el segundo y finalmente un tercer caso con ambas metodologías.

6. Aun así aplicándose el algoritmo EM y la Imputación Múltiple a un conjunto de datos perdidos completamente al azar MCAR, se obtuvieron estimaciones cercanas a los obtenidos en la muestra original, demostrando que hasta con mecanismos de pérdida no considerados en su metodología la eficiencia de ambas metodologías ha sido óptima.

7. Se comprobó en tablas comparativas la precisión y cercanía de las estimaciones obtenidas con el algoritmo EM en la Imputación Múltiple y las obtenidas de la muestra original (sin datos faltantes). Dichas estimaciones en el caso del método Listwise Deletion se mostraron no tan precisos y en ciertos casos con valores subestimados con respecto a las estimaciones de la muestra original.

8. Por lo tanto, es importante recalcar que los métodos tradicionales de imputación presentan ciertas deficiencias y las metodologías como el algoritmo EM y la Imputación Múltiple que tienen base teórica respectivamente en la estimación por máxima verosimilitud y en el enfoque bayesiano han mostrado tener mejor precisión de estimación en muestras con datos faltantes.

REFERENCIAS BIBLIOGRÁFICAS

Alfaro, R., & Fuenzalida, M. (2009). Imputación Múltiple en Encuestas Microeconómicas. *Cuadernos de economía*, *46*(134), 273-288.

Anderson, T.W. (1957). Maximum likelihood estimates for the multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, *52*, 200–203.

Cañizares, M., Barroso, I., & Alfonso, K. (2004). Datos incompletos: una mirada crítica para su manejo en estudios sanitarios. *Gaceta Sanitaria*, 18(1), 58-63.

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1997). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39-138.

Díaz, R. P. Casos con Datos Faltantes:¿ Qué Hacer con Ellos?.

Enders, C. K. (2010). *Applied missing data analysis*. The Guilford Press.

Feres, J. C. (1998). Falta de respuesta a las preguntas sobre el ingreso. Su magnitud y efectos en las Encuestas de Hogares en América Latina. *Documento presentado en el 2º Taller Regional de Mecovi, Buenos Aires.*

Galván, M., & Medina, F. (2007). *Imputación de Datos: Teoría y Práctica*. UN.

García, J. G., Albaladejo, J. P., & Fernández, J. A. M. (2006). Métodos de inferencia estadística con datos faltantes: estudio de simulación sobre los efectos en las estimaciones. *Estadística española*, *48*(162), 241-270.

Harel, O., & Zhou, X. H. (2006). Multiple imputation-Review of theory, implementation and software. Stat *Med*., *26(16)*, 3057–3077.

Horton, N.J. & Lipsitz, S.R. (2001). Multiple imputation in practice: Comparison of software package for regression models with missing variables. *The American Statistician, 55,* 244–254.

Li, K.H. (1988). Imputation using Markov chains. *Journal of Statistical Computation and Simulation*, *30*, 57–79.

Little, R. J., & Rubin, D. B. (1987). *Statistical analysis with missing data* (Vol. 539). New York: Wiley.

Méndez Martín, J. M. (2008). Imputation in the survey on living conditions. *BEIO, Boletín de Estadística e Investigación Operativa*, *24*(1), 25-28.

Meng, X. L., & Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2), 267-278.

Peña, D. (2002). *Análisis de datos multivariantes* (Vol. 24). Madrid: McGraw-Hill.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical methods in medical research*, *8*(1), 3-15.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate behavioral research*, 33(4), 545-571.

Scheffer, J. (2002). Dealing with missing data. *Research letters in the information and mathematical sciences*, *3*(1), 153-160.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *the Annals of Statistics*, 1701-1728.

Useche Castro, L. M., & Mesa Avila, D. M. (2011). Una introducción a la imputación de valores perdidos. *Terra. Nueva Etapa*, 22(31).

Yuan, Y. C. (2010). Multiple imputation for missing data: Concepts and new development (Version 9.0). *SAS Institute Inc, Rockville, MD.*

Yupanqui, R. M. (2005). Introducción a la estadística bayesiana. (Tesis de Licenciatura publicada). Lima, Perú: Universidad Nacional Mayor de San Marcos.

Zhang, P. (2003). Multiple imputation: theory and method. *International Statistical Review*, *71*(3), 581-592.

ANEXOS

USO DEL PROGRAMA MATLAB EN EL DESARROLLO DE TEMA DE INVESTIGACIÓN

SOBRE MATLAB

Matlab es una herramienta de software matemático disponible para las plataformas Unix, Windows, Mac OS X y GNU/Linux. Entre las prestaciones de Matlab están la manipulación de matrices, la representación de datos y funciones, la implementación de algoritmos, la creación de interfaces de usuario (GUI) y la comunicación con programas en otros lenguajes.

Para obtener información sobre el programa así como una versión de prueba se encuentra disponible http://www.mathworks.com/products/matlab/

PROGRAMA MATLAB APLICADO A LAS METODOLOGÍAS DEL TRABAJO DE INVESTIGACIÓN

Los algoritmos son un conjunto de instrucciones o reglas ordenadas y finitas que te permiten ejecutar mediante pasos sucesivos alguna actividad, para el presente trabajo se realizaron dos algoritmos para los métodos de estimación por máxima verosimilitud y la imputación múltiple.

Este proceso se realizó con la ayuda de pseudocódigos de las metodologías que están descritas en el capítulo III del presente trabajo de tesis.

CÓDIGOS EN MATLAB DE LOS PROCEDIMIENTOS DE IMPUTACIÓN

Para ejecutar la estimación por máxima verosimilitud se trabajó con el algoritmo EM, método por el cual a través de valores de parámetros iniciales provenientes de la matriz de datos observada se obtienen los valores que sustituyen a los faltantes, este procedimiento es denominado Paso E. Luego de obtener una matriz de datos completa se continúa con el paso M, donde se obtendrán los nuevos parámetros estimados con la nueva matriz de datos completa; ambos pasos se ejecutarán un número de veces convenientes hasta que las diferencias entre parámetros estimados en cada paso sean mínimas. A continuación presentamos el código utilizado en el programa Matlab para ejecutar la estimación por Máxima Verosimilitud utilizando el Algoritmo EM.

ALGORITMO EM

```
%ALGORITMO EM (EXPECTATION MAXIMIZATION)
clear;
clc;
X=xlsread('DataP.xls');%SE CARGA LOS DATOS COMPLETOS DE ARCHIVO DataP.xls
n=size(X);

for j=1:n(2)
    suma=0;
    cont=0;
for i=1:n(1)
    if X(i,j)>-99
        suma=suma+X(i,j);
        cont=cont+1;
    end
end
mediat0(j)=suma/cont;

end
mediat0;
cont=0;
```

```
for i=1:n(1)
    if min(X(i,:))>-99
        cont=cont+1;
    XOBS(cont,:)=X(i,:);
    end
end
XOBS;

%Se especifican los puntos iniciales, que es la vector de media y matriz
%de covarianza de la matriz de datos observada.
a1=input('LOS PARÁMETROS INICIALES SON:');
mediat0
Vt0=(cont-1)*cov(XOBS)/cont

%PASO E: Calcula y reemplaza en los valores faltantes la esperanza de la
función de verosimilitud de los datos completos con respecto a la
distribución de los datos faltantes utilizando los valores de vector de
medias (mediat0) y matriz de covarianza (Vt0) iniciales definidas.

for t=1:50
    matriz0=zeros(n(2));
for i=1:n(1)
    if min(X(i,:))==-99;
        cont1=0;
        cont2=0;
        for j=1:n(2)
            if X(i,j)==-99;
                cont1=cont1+1;
                indices1(cont1)=j;
            else
                cont2=cont2+1;
                indices2(cont2)=j;
            end
        end

        media1=mediat0(indices1);
        media2=mediat0(indices2);
        VORD=zeros(n(2));
        indices=[indices1 indices2];

        for jf=1:n(2)
            for jc=1:n(2)
                VORD(indices(jf),indices(jc))=Vt0(jf,jc);
            end
        end

     V11=VORD(1:length(indices1),1:length(indices1));
     V12=VORD(1:length(indices1),(length(indices1)+1):n(2));
     V22=VORD((length(indices1)+1):n(2),(length(indices1)+1):n(2));
     x2=X(i,indices2);
     Ex1idx2i=media1'+V12*inv(V22)*(x2-media2)';
```

```
    cont=0;
    for j=1:n(2)
        if X(i,j)==-99
            cont=cont+1;
            XIMPUT(i,j)=Ex1idx2i(cont);
           else
            XIMPUT(i,j)=X(i,j);
        end
    end

    Ex1ix1i=V11-V12*inv(V22)*V12'+Ex1idx2i'*Ex1idx2i;
    Ex2ix2i=x2'*x2;
    Ex1ix2i=Ex1idx2i'*x2;
    SCDES=[Ex1ix1i Ex1ix2i; Ex1ix2i' Ex2ix2i];
    SC=zeros(n(2));

     for jf=1:n(2)
            for jc=1:n(2)
                SC(indices(jf),indices(jc))=SCDES(jf,jc);
            end
     end

    else
        XIMPUT(i,:)=X(i,:);
        SC=X(i,:)'*X(i,:);
    end
    matriz0=matriz0+SC;
   end
t;
XIMPUT;

%Paso M: Calcula los nuevos estimadores de la nueva matriz de datos
completa(XIMPUT) con los valores reemplazados por las estimaciones
obtenidos en el paso E precedente.

mediat0=mean(XIMPUT);
Vt0=matriz0/n(1)-mediat0'*mediat0;
d=diag(Vt0)';
Res(:,t)=[t mediat0(1,1:j) d(1,1:j) Vt0(1,j-1:j) Vt0(j-1,j)];
end

% Los Resultados se muestran en la siguiente matriz

a2=input('LAS ESTIMACIONES OBTENIDAS EN CADA ITERACIÓN:');
Resultados=Res'
```

A continuación presentamos el Código utilizado para el método de imputación múltiple de datos faltantes, este procedimiento es un proceso que imputa valores faltantes seleccionando valores que tienen como principal componente un valor aleatorio del dato imputado, estas realizaciones se obtienen a través de la caracterización de la distribución conjunta de los datos, que por lo general se asume normal.

ALGORITMO DE IMPUTACIÓN MÚLTIPLE

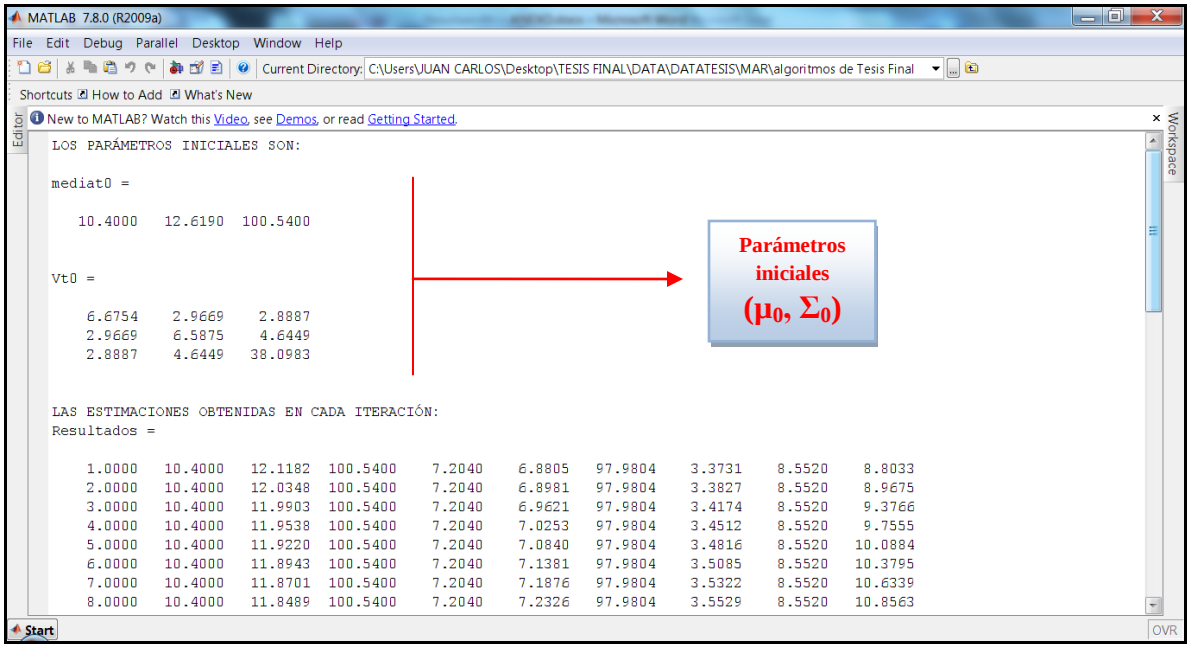
```
%IMPUTACIÓN MÚLTIPLE DE DATOS (MULTIPLE IMPUTATION)
clear;
clc;

X=xlsread('DataP.xls');%SE CARGA LOS DATOS COMPLETOS DE ARCHIVO Data.xls
n=size(X);

%Se especifican los puntos iniciales, que particularmente son
%los puntos que se obtuvieron en el proceso de estimación
%del algoritmo EM.

%Puntos iniciales obtenidos del algoritmo EM
a1=input('LOS PARÁMETROS INICIALES SON:');
mediat0=[10.4000   11.7022  100.5400 ]

Vt0=[7.2040    3.6970    8.5520
     3.6970    7.5893   12.3949
     8.5520   12.3949   97.9804]

%Paso I: Se obtendrán los valores imputados que depende
%del vector de medias y matriz de covarianzas iniciales

TETHA0=zeros(1,n(2));
VW0=zeros(n(2));
VB0=zeros(n(2));

a2=input('SE INICIAN LAS IMPUTACIONES ... ');
```

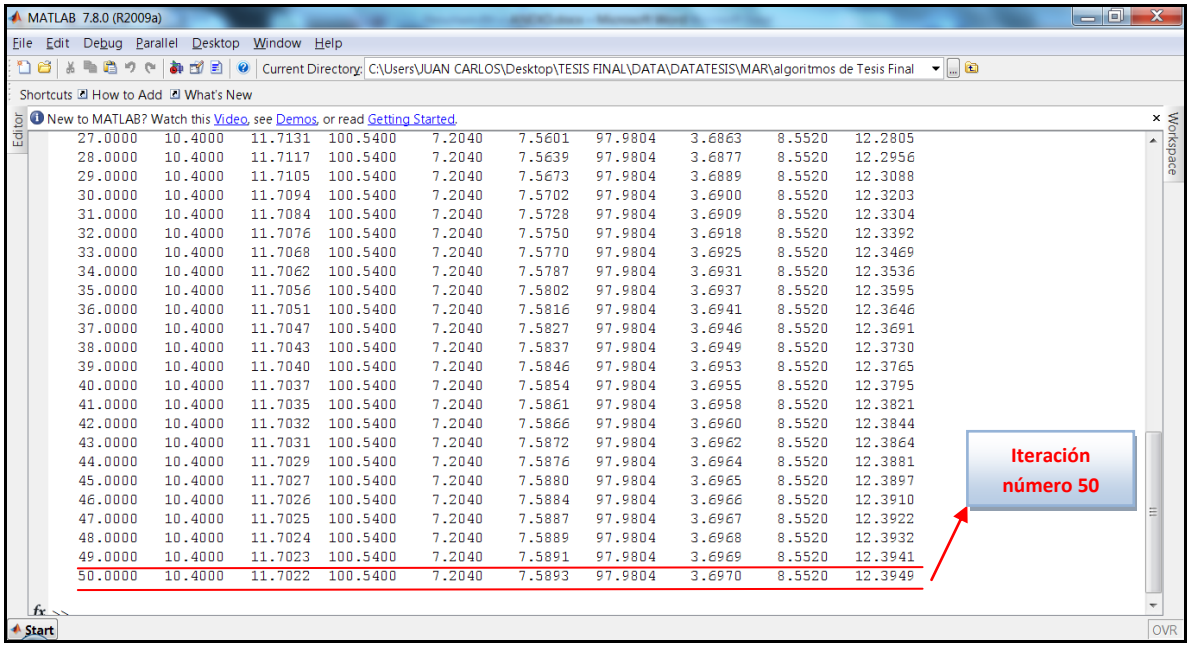
```
for t=1:10
for i=1:n(1)
    if min(X(i,:))==-99
            cont1=0;
            cont2=0;
        for j=1:n(2)
            if X(i,j)==-99
                cont1=cont1+1;
                indices1(cont1)=j;
            else
                cont2=cont2+1;
                indices2(cont2)=j;
            end
        end
        U1=mediat0(indices1);
        U2=mediat0(indices2);
        VORD=zeros(n(2));
        indices=[indices1 indices2];
        for jf=1:n(2)
            for jc=1:n(2)
                VORD(indices(jf),indices(jc))=Vt0(jf,jc);
            end
        end

     V11=VORD(1:length(indices1),1:length(indices1));
     V12=VORD(1:length(indices1),(length(indices1)+1):n(2));
     V22=VORD((length(indices1)+1):n(2),(length(indices1)+1):n(2));
     U1dU2=U1+V12*inv(V22)*(X(i,indices2)-U2)';
     VU1dU2=V11-V12*inv(V22)*V12';
     vimput=mvnrnd(U1dU2,VU1dU2);
     cont3=0;
        for j2=1:n(2)
            if X(i,j2)==-99
                cont3=cont3+1;
                XIMPUT(i,j2)=vimput(cont3);
else
                XIMPUT(i,j2)=X(i,j2);
            end
        end
            else
                XIMPUT(i,:)=X(i,:);
    end

end
t

mediat0=mean(XIMPUT)
TETHA0=TETHA0+mediat0(1,1:n(2));
```

```
Vt0=(n(1)-1)*cov(XIMPUT)/n(1)
VT(:,:,t)=Vt0;
VW0=VW0+Vt0(1:n(2),1:n(2));%PARA OBTENER LAS VARIANZAS DENTRO DE LA
IMPUTACIÓN

d=diag(Vt0)';
Res(:,t)=[t mediat0(1,1:j) d(1,1:j) Vt0(1,j-1:j) Vt0(j-1,j)];
end

%OBTENCIÓN DE ESTIMACIONES
a3=input('LAS ESTIMACIONES EN CADA IMPUTACIÓN:');
Resultados=Res'

a4=input('LAS ESTIMACIÓN GLOBAL DE LA MEDIA:');
TETHA=TETHA0/t

a5=input('LA VARIANZA DENTRO DE LA IMPUTACIÓN ES:');
VW=VW0/t

a6=input('LA VARIANZA ENTRE IMPUTACIONES ES:');
for t=1:10
VB1=(VT(:,:,t)-VW)*(VT(:,:,t)-VW)';
VB0=VB0+VB1;
end
VB=VB0/(t-1)

a7=input('LA VARIANZA TOTAL OBTENIDA ES:');
T=VW+((1+(1/t))*VB)

a8=input('LOS PARÁMETROS ESTIMADOS RESPECTIVOS SON:');
TETHA
T
```

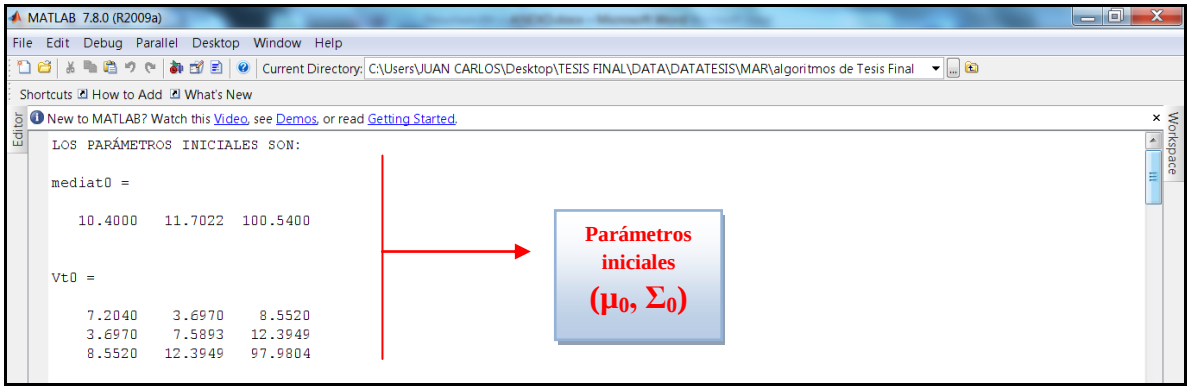
Los resultados de las corridas de ambos algoritmos se muestran a continuación:

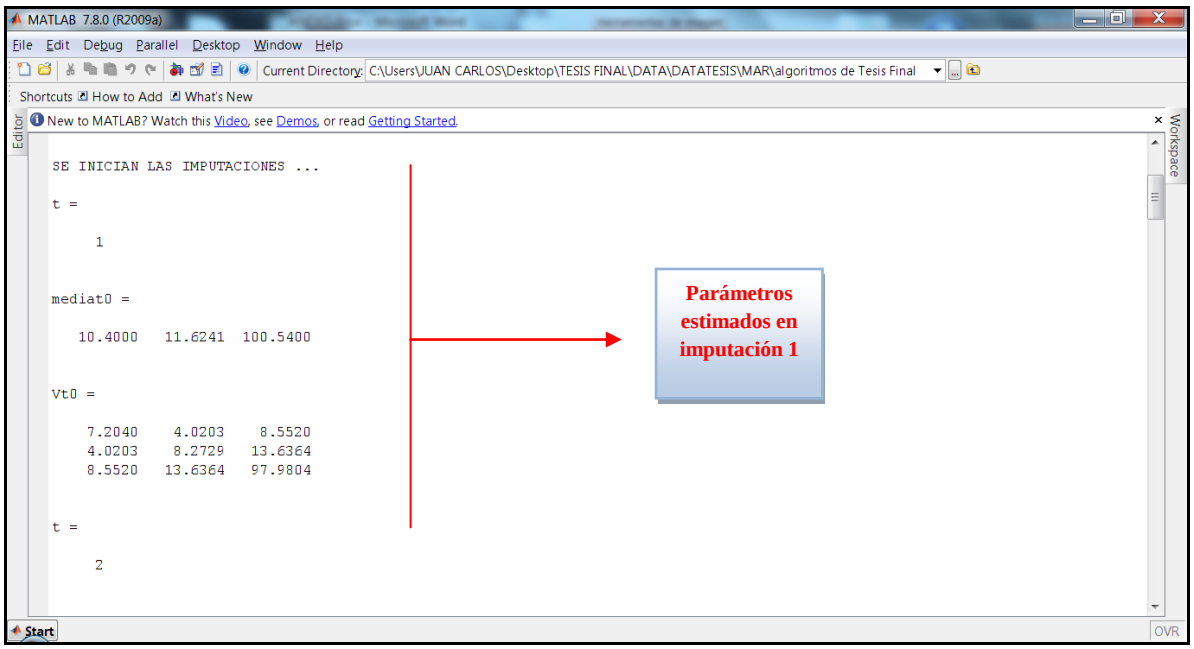
Primero presentamos los resultados obtenidos con el algoritmo EM, primero nos presentan los parámetros iniciales (μ_0, Σ_0) y posteriormente las estimaciones obtenidas corresponden a $(t, \hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3, \hat{s}_1^2, \hat{s}_2^2, \hat{s}_3^2, \hat{s}_{12}, \hat{s}_{13}, \hat{s}_{23})$ en las 50 iteraciones.



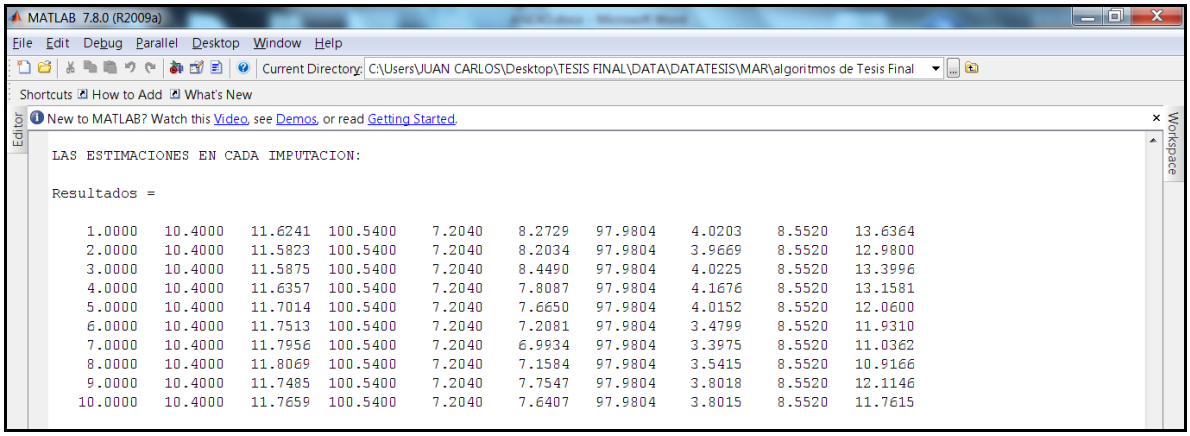


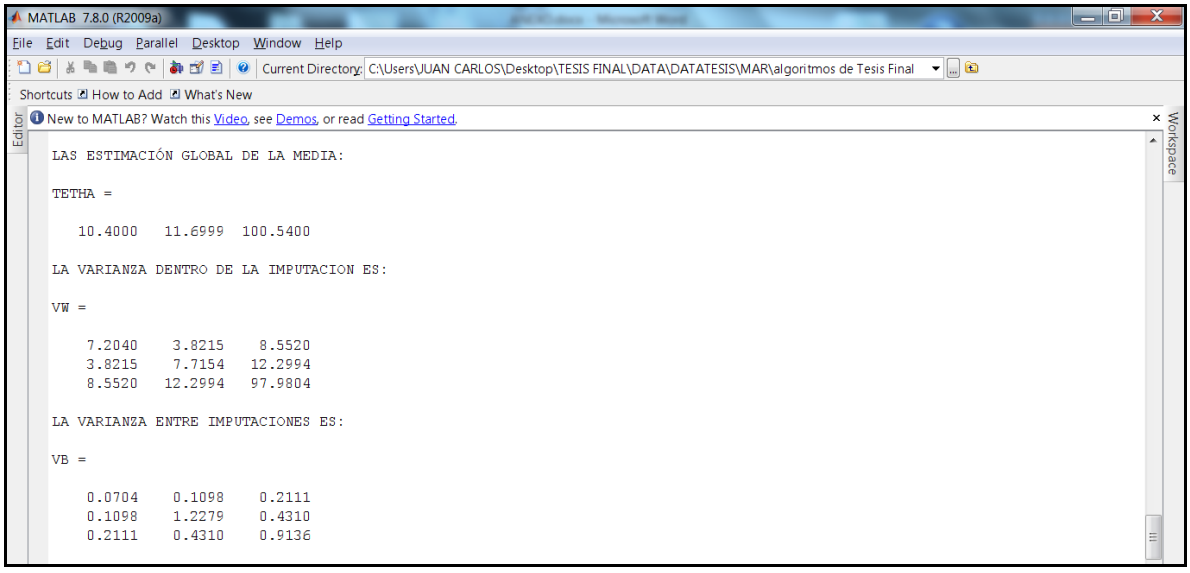
Posteriormente presentamos los resultados obtenidos con el algoritmo IM, establecemos los parámetros iniciales (μ_0, Σ_0) que en el presente caso son las estimaciones obtenidas a través de algoritmo EM en la iteración número 50, posteriormente se especifica t; que es el número de conjuntos de datos que se van a generar (para el presente caso se generan $t = 10$).

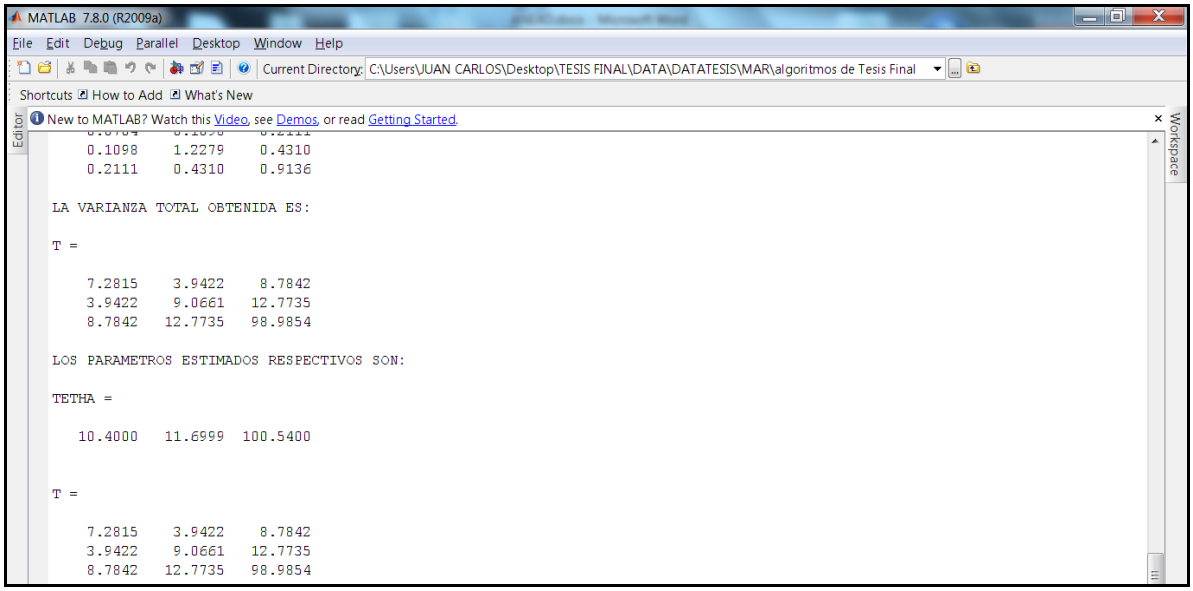




De manera resumida se presenta los resultados del proceso de Imputación del método IM, donde se muestra las $t = 10$ estimaciones obtenidas $(t, \hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3, \hat{s}_1^2, \hat{s}_2^2, \hat{s}_3^2, \hat{s}_{12}, \hat{s}_{13}, \hat{s}_{23})$ de los m conjuntos de datos que se imputaron para posteriormente seguir con el proceso de análisis y combinación del método de imputación múltiple.


MATLAB 7.8.0 (R2009a)
File Edit Debug Parallel Desktop Window Help
Current Directory: C:\Users\JUAN CARLOS\Desktop\TESIS FINAL\DATA\DATATESIS\MAR\algoritmos de Tesis Final
Shortcuts How to Add What's New
New to MATLAB? Watch this Video, see Demos, or read Getting Started.
Editor
Workspace
LAS ESTIMACIONES EN CADA IMPUTACION:
Resultados =
1.0000 10.4000 11.6241 100.5400 7.2040 8.2729 97.9804 4.0203 8.5520 13.6364
2.0000 10.4000 11.5823 100.5400 7.2040 8.2034 97.9804 3.9669 8.5520 12.9800
3.0000 10.4000 11.5875 100.5400 7.2040 8.4490 97.9804 4.0225 8.5520 13.3996
4.0000 10.4000 11.6357 100.5400 7.2040 7.8087 97.9804 4.1676 8.5520 13.1581
5.0000 10.4000 11.7014 100.5400 7.2040 7.6650 97.9804 4.0152 8.5520 12.0600
6.0000 10.4000 11.7513 100.5400 7.2040 7.2081 97.9804 3.4799 8.5520 11.9310
7.0000 10.4000 11.7956 100.5400 7.2040 6.9934 97.9804 3.3975 8.5520 11.0362
8.0000 10.4000 11.8069 100.5400 7.2040 7.1584 97.9804 3.5415 8.5520 10.9166
9.0000 10.4000 11.7485 100.5400 7.2040 7.7547 97.9804 3.8018 8.5520 12.1146
10.0000 10.4000 11.7659 100.5400 7.2040 7.6407 97.9804 3.8015 8.5520 11.7615


MATLAB 7.8.0 (R2009a)
File Edit Debug Parallel Desktop Window Help
Current Directory: C:\Users\JUAN CARLOS\Desktop\TESIS FINAL\DATA\DATATESIS\MAR\algoritmos de Tesis Final
Shortcuts How to Add What's New
New to MATLAB? Watch this Video, see Demos, or read Getting Started.
Editor
Workspace
LAS ESTIMACIÓN GLOBAL DE LA MEDIA:
TETHA =
10.4000 11.6999 100.5400
LA VARIANZA DENTRO DE LA IMPUTACION ES:
VW =
7.2040 3.8215 8.5520
3.8215 7.7154 12.2994
8.5520 12.2994 97.9804
LA VARIANZA ENTRE IMPUTACIONES ES:
VB =
0.0704 0.1098 0.2111
0.1098 1.2279 0.4310
0.2111 0.4310 0.9136


MATLAB 7.8.0 (R2009a)
File Edit Debug Parallel Desktop Window Help
Current Directory: C:\Users\JUAN CARLOS\Desktop\TESIS FINAL\DATA\DATATESIS\MAR\algoritmos de Tesis Final
Shortcuts How to Add What's New
New to MATLAB? Watch this Video, see Demos, or read Getting Started.
Editor
Workspace
0.1098 1.2279 0.4310
0.2111 0.4310 0.9136
LA VARIANZA TOTAL OBTENIDA ES:
T =
7.2815 3.9422 8.7842
3.9422 9.0661 12.7735
8.7842 12.7735 98.9854
LOS PARAMETROS ESTIMADOS RESPECTIVOS SON:
TETHA =
10.4000 11.6999 100.5400
T =
7.2815 3.9422 8.7842
3.9422 9.0661 12.7735
8.7842 12.7735 98.9854